Bioinformatics to improve shotgun proteomics pipeline. Towards an efficient MALDI-MS pipeline

Gandhi, Tejas Paresh

BIOINFORMATICS TO IMPROVE SHOTGUN PROTEOMICS ANALYSES

TOWARDS AN EFFICIENT MALDI-MS PIPELINE

Tejas Gandhi

Cover: Eduard M Perez (info@dsign-ed.com)

Inside artwork: Tejas Gandhi

Printed and bound in The Netherlands by Ipskamp Drukkers

This PhD study was carried out in the Biochemistry Department of the *Groningen Biomolecular Sciences and Biotechnology Institute (GBB)* at the *University of Groningen* and was supported by *The Netherlands Bioinformatics Center (NBIC)*.

rijksuniversiteit
groningen

BIOINFORMATICS TO IMPROVE SHOTGUN PROTEOMICS ANALYSES

TOWARDS AN EFFICIENT MALDI-MS PIPELINE

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
maandag 27 juni 2011
om 13:15 uur

door

Tejas Paresh Gandhi

geboren op 5 juli 1979
te Ahmedabad, India

Promotores:	Prof. dr. B. Poolman Prof. dr. R. Breitling
Copromotor:	Dr. H. P. Permentier
Beoordelingscommissie:	Prof. dr. R. Bischoff Prof. dr. A. J. Heck Prof. dr. O. P. Kuipers

ISBN: 978-90-367-4955-8

to my parents…

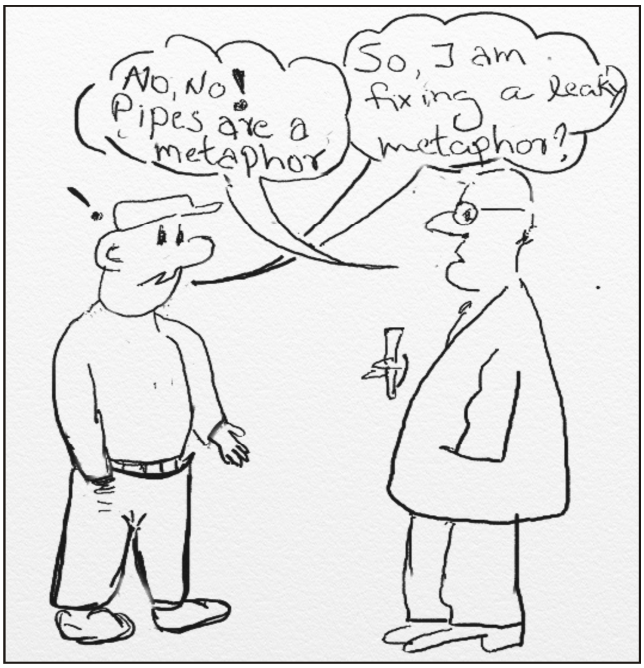
CONTENTS

Prologue

The path in front of me was straight as an arrow, as if mocking me. Odd for an autumn day in Groningen, but the leaves were hardly stirring. I nervously glanced around to find no one else in the vicinity. The conditions were ripe, but a recent bruise reminded me of the price of failure. As I hesitated, dark thoughts entered my mind. I suddenly questioned the sanity of continuing to stick to an endeavor that had shown no positive result after nearly two years. I argued, "There are other things I can achieve in life! Like… I could take up juggling." Sagely advice of *being patient* and *not giving up* seemed like a cruel joke. But, soon the stubbornness kicked in, as I whispered a quote from Frank Herbert's Dune: *Fear is the mind-killer… I will face my fear*. Taking a deep breath, I straightened my back, narrowed my eyes and focused on the passage in front of me. My ears tuned into the soft, rapid, clicking noise of the derailleur gears as I pedaled forward with an increasing gusto. I was ready.

Letting go of one of the handlebars, I felt the balance through my hips. A moment later, the second handlebar too, as I pedaled faster seeking to exploit the momentum. The moment of reckoning: I negotiated the first ten meters perfectly, but I knew from my past failures that the challenge was in the next ten. Then, the slightest inclination as my front wheel started steering towards its right. Left uncorrected, it would prove to be the cause of yet another sore disappointment. Fending off feelings of despair, I used my lower body to recalibrate. My legs shook and the balance, at best, seemed precarious. But, it was working. I burst out in laughter as I had finally managed to ride my *Batavus* without the use of my hands!

Chapter I



Introduction to Shotgun Proteomics

1. Introduction

In computer science, a pipeline is defined as a set of serially connected data processing elements, such that the output of one element is the input of the next one. This definition, however, can be easily applied to any other field. For example, switch the *data processing elements* with enzymes in biochemistry and you end up with a pipeline describing a metabolic pathway such as glycolysis. The key notion to remember, when dealing with a pipeline, is that of blackboxing, which means that the elements of a pipeline can be abstracted. This is crucial in the light of the fact that each element within a pipeline can be a pipeline in its own right. So, a pipeline serves as a tool for abstracting often confounding details into the realm of the manageable. This type of abstraction allows treating its individual elements independently, but improvement in the efficiency of a *part* reflects in the enhanced performance of the *whole*. The body of this work grapples with the task of improving the performance of a pipeline central to the large-scale study of proteins, namely shotgun proteomics, by improving different parts of it.

1.1 Shotgun proteomics: a really, really big jigsaw puzzle

Proteins play an overwhelmingly dominant role in living organisms as the work-horses of cells. The term proteomics, coined to serve as an analogy to genomics, is often defined as the comprehensive, quantitative study of protein expression and its changes under the influence of biological perturbations.[1,2] The goal of a typical proteomics experiment is to juxtapose the set of expressed proteins from a living organism under different conditions such as temperature, mutation, nutrient availability, disease vs. healthy. The idea is to understand a biological system by looking at the differences between two or more states, in case of proteomics in the content, state (qualitative), and levels (quantitative) of proteins. Obviously, before any thorough quantitative analysis of differences can be undertaken, unknown proteins expressed in a biological sample must first be identified.

The term 'shotgun proteomics' refers to the *pipeline* used for identifying proteins in a complex protein mixture. Sample preparation in shotgun proteomics is relatively simple; as the name implies no separation on the protein level is performed, but instead, the entire complex sample of protein is digested into an even more complex mixture of peptides. The analytical part of the pipeline typically encompasses liquid chromatography (LC) based separation of peptides coupled with (tandem) mass

spectrometry (MS(/MS)).[3,4,5] Although the analytical LC and MS methods may differ across various shotgun proteomics strategies (e.g. on-line or off-line fractionation of peptides), there are several common steps as illustrated in Figure 1. The general strategy is conceptually analogous to solving a jigsaw puzzle, with two additional challenges: 1) it can involve hundreds of puzzles (proteins) with their pieces (peptides) mixed together, and 2) the pieces are invisible to the naked eye. The latter is addressed using mass spectrometry, an analytical technique that measures compounds very accurately by their mass-to-charge ratio (m/z), and that in addition can specifically fragment compounds to obtain structural information.

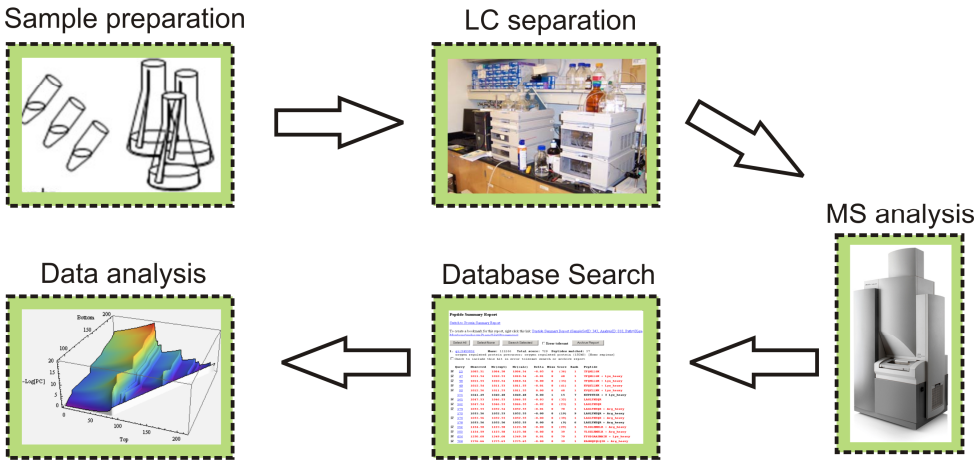


Figure 1. A typical 2D-LC-MS proteomics pipeline

In this pipeline, proteins of a biological sample are typically digested with trypsin (Sample preparation). The resulting peptides are separated by liquid chromatography (LC) and, upon elution, ionized in the mass spectrometer (MS) for characterization by mass separation of peptide ions. The mass spectrometer can subsequently select peptide ions for fragmentation (MS/MS) to yield fragment ion masses that are then used to identify the peptide and the corresponding proteins, using database search engines. Since MS data contains both qualitative (ion mass) and quantitative (ion intensity) information, quantification can take place at the same time as identification, which requires additional data analysis.

1.2 The mass spectrometry approach: solving a puzzle within a puzzle

In order to be amenable to MS analysis, the unknown proteins in a sample need to be first digested into peptides, usually done using a protease such as trypsin. The primary unknown factor in proteomics is the identity of the proteins in the sample (puzzles), whereas the known feature is the database of protein sequences (all possible solutions). The corresponding *in-silico* digested theoretical peptides (jigsaw pieces) can easily be derived from the protein database by applying the rules of enzyme specificity. For example, trypsin is known to exclusively cut proteins after lysine (Lys) and arginine (Arg) residues.[7]

Protein digestion is followed by an initial sorting similar to sorting of puzzle pieces based on features such as shape or colour. Here, the peptides are separated using a one- or multi-dimensional fractionation, typically liquid chromatography (LC). Chromatographic separation is achieved based on physicochemical features such as peptide charge (in case of strong cation exchange LC) or hydrophobicity (reversed phase LC). Each chromatographic peak or fraction of peptides is then analyzed by MS. Peptides are ionized in the ionization source and peptide ions are separated according to their mass-to-charge ratio (m/z). The mass of a peptide ion can be predicted from its composition of amino residues and ultimately its elemental composition, whereas its charge is dependent on both the chemical properties of the peptides (basicity) and the ionization method. The knowledge of experimental m/z values is used to identify proteins, taking advantage of the known database of theoretical peptides by computer search engines such as Mascot and Sequest.[7,8] The protein and peptide assignments of these search engines are qualified by some type of a score indicating the associated confidence level. Levels of identification confidence are generally expressed as percentage values, and threshold values of between 95 and 99 % are often applied for reporting identified proteins.

Considering the complexity of most proteomics samples, it is not difficult to imagine that multiple jigsaw pieces are (almost) identical to each other but stemming from different puzzles, i.e. their m/z values are so close (or even identical) that they cannot be discriminated with the current MS instruments.[9] This is a fairly common phenomenon amongst peptides in a complex mixture of proteins. In order to resolve this problem, the peptides are typically broken into fragment ions and analyzed by a second round of MS. This works because peptides with very similar or identical mass will often lead to very different fragments in MS/MS mode, due to their different amino

acid sequence.[10,11] In addition, many of the rules of fragmentation are known (Figure 2) and therefore, just as for peptide digestion, *in silico* prediction of fragment masses can be used for comparison and identification by database search programs such as Mascot. In this manner, a set of jigsaw subpieces (fragment ions) is first assembled into a jigsaw piece (peptide), which ultimately leads to the final answer of protein identification (complete puzzle).

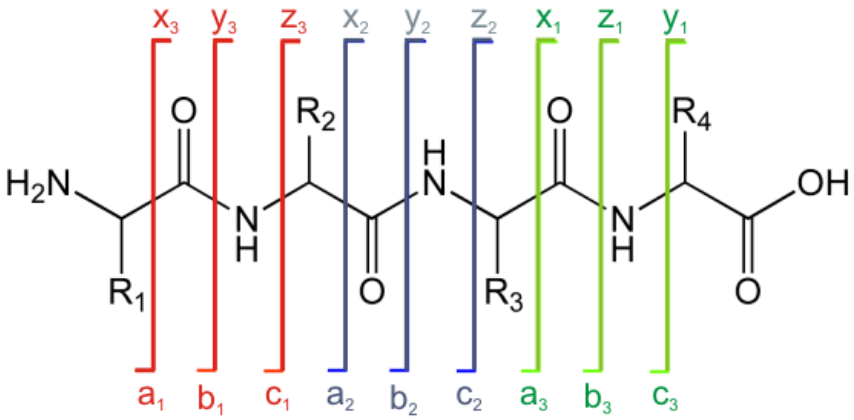


Figure 2. Notation schema of peptide backbone fragmentation

Peptide fragments are divided into two groups: 1) a, b, c and 2) x, y, z, based on whether the charge is retained on the N-terminus or C-terminus, respectively.[12] The subscript indicates the number of amino acid residues in the fragment. Typically, *b*- and *y*-ions are the most commonly occurring fragments with soft-ionization methods such as ESI and MALDI.

1.3 MALDI-based proteomic analysis

As mentioned in section 1.1, there are a variety of options available for an analytical setup for shotgun proteomics, each with its strengths and weaknesses. As is often the case, software methodological developments are highly dependent on the setup and may produce high-quality results only within the context of their development.[13] All the experiments presented in this work stem from a two-dimensional liquid chromatography (2D-LC) system coupled to a *Matrix-Assisted Laser Desorption Ionization Time of Flight/Time of Flight* (MALDI-TOF/TOF) instrument. While a comparison of various setups is outside the scope of this work, the LC-MALDI-

TOF/TOF system has several specific characteristics worth mentioning. Being an off-line system, meaning that the LC separation and subsequent MS steps are separated, it facilitates a data-driven analysis.[14,15] Also, due to peculiarities of the ionization process, MALDI spectra are overwhelmingly dominated by singly-charged peptide ions irrespective of amino acid composition, the MS spectra and the subsequent identification process are simplified.[16]

1.4 iTRAQ-based protein quantification

LC-MS/MS is widely used, not only for identification, but also for quantification of proteins, using either isotopically labelled peptides or label-free quantification.[17] Most data sets analyzed in this thesis were obtained from complex proteome samples derivatized with the isotopic label iTRAQ (isobaric Tags of Relative and Absolute Quantification) reagents in order to derive relative quantitative information.[18,19] Isotopic labelling is considered to be more reliable for relative quantitation than label-free methods, since in the former the samples to be compared are mixed together and analyzed simultaneously and therefore do not suffer from inaccuracies due to technical fluctuations between analysis runs.

The iTRAQ reagents are reactive towards amine groups and therefore lead to chemical modification of the N-terminal peptide amine and lysine residues, but some cross-reactivity with tyrosine residues can be observed. Two types of iTRAQ reagents are commercially available, namely 4-plex and 8-plex, which can be used to label and differentially quantify four or eight different samples, respectively. The reagent itself (Figure 3) consists of an amine-reactive group (N-hydroxysuccinimide-ester), a reporter group (for quantitation), and a balance group (which keeps the total reagent mass the same). The total mass of the reporter plus balance group of each 4-plex reagent is 145 Da, whereas that of the 8-plex reagent is 305 Da. However, the masses of reporter groups are different for each reagent within the 4-plex and 8-plex sets: 114 through 117 Da for 4-plex and 113 through 119 plus 121 Da for 8-plex. The masses of the balance group are different accordingly. Different masses of reporter and balance are achieved by incorporating different combinations of heavy and light isotopes of their constituent atoms, specifically C, N, and presumably O (the chemical structure of the balance group is C=O in 4-plex iTRAQ, but it is unpublished for 8-plex iTRAQ). Furthermore, the overall charge and hydrophobicity of the peptides with iTRAQ modifications remains the same such that they co-elute in all LC separations.

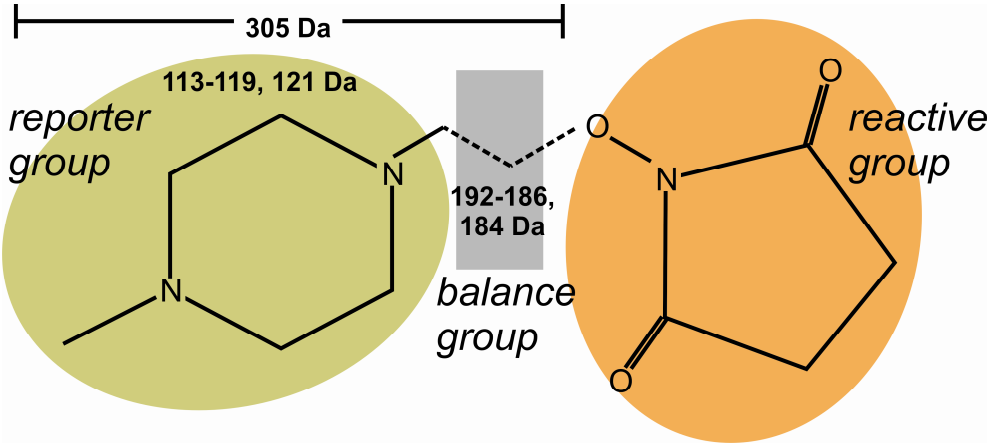


Figure 3. Schematic structure of 8-plex iTRAQ reagent
The 8-plex iTRAQ reagent is compromised of three different components: reporter, balance, and reactive group, and comes in a set of 8 different forms. While the reactive group is the same across the set, the reporter group can have 8 different masses: 113-119, and 121 Da. The total mass is conserved by the balance group which adjusts accordingly: 192-186, and 184 Da.

This arrangement of the iTRAQ label means that two peptides with the same amino acid sequence but tagged by two different iTRAQ reagents will be detected in the same LC fraction during the separation and at the same mass (m/z) in the MS mode. However, upon fragmentation in the MS/MS mode the different reporter groups from each peptide separate from the balance group and will be detected as a peak at different m/z values. The relative intensities of the peaks can be used as a measure for the relative abundance of the two differently tagged peptides, and by extension, of the protein from which they are derived. In this way, peptides from e.g. a treated and control sample, tagged with a different iTRAQ reagent, can be identified and compared in a quantitative manner at the same time.

2. Challenges in the shotgun proteomics pipeline

Despite the successful identification of increasingly large numbers of proteins by mass spectrometric techniques, owing to improvements in MS instrument sensitivity and accuracy, inherent methodological constraints and the underlying complexity of a typical proteome still prevent truly comprehensive coverage.[20,21] For instance, even with

the relatively small yeast proteome a comprehensive coverage has not been achieved.[21,22,23] Additionally, in recent years, the attention of the proteomics community has started to shift from ever longer lists of identified proteins towards their accurate quantification. Relative quantitation methods suffer from lack of dynamic range and reproducibility issues, in case of isotopically labeled samples, and from availability of appropriate standards for label-free quantification. All quantitative methods have in common the need for accurate and statistically correct data analysis. In practice, changes in protein concentrations in proteome samples are rather small, as they are in many cases controlled within a narrow range (homeostasis) by complex cellular regulation processes. Measurement of statistically significant changes of these small changes is thus challenging with current quantitative shotgun proteomics methods.

2.1 The membrane proteome: invariably a problem-child

One of the factors contributing to the difficulty of achieving full proteome coverage is the large group of membrane proteins. Owing to the challenge of isolating them during sample preparation and the large hydrophobic stretches which are often inaccessible for proteolytic or chemical digestion, membrane proteins are prone to be 'invisible' to MS (Figure 4). Generally, membrane proteins are identified predominantly by the peptides arising from their soluble domains. This means that membrane proteins lacking large soluble domains tend to be left unidentified. Sample preparation is likely to be the step where the fate of these types of proteins is decided. Use of proteases other than trypsin has been presented as a potential solution for increasing their visibility in the past.[25,26] Cleavage specificity for hydrophobic amino acid residues, such as chymotrypsin, which cleaves after bulky, aromatic residues, helps in producing peptides from transmembrane segments of masses amenable to MS analysis (Figure 4). However, accessibility of the transmembrane segment for the protease and subsequent recovery of smaller, but still hydrophobic, peptides still lead to low sequence coverage of many membrane proteins in shotgun proteomics. In this work, we do not experiment with proteases other than trypsin, but arguably improvements in other steps of the proteomics pipeline will also lead to an improved identification and quantification of membrane proteins via their soluble domains.

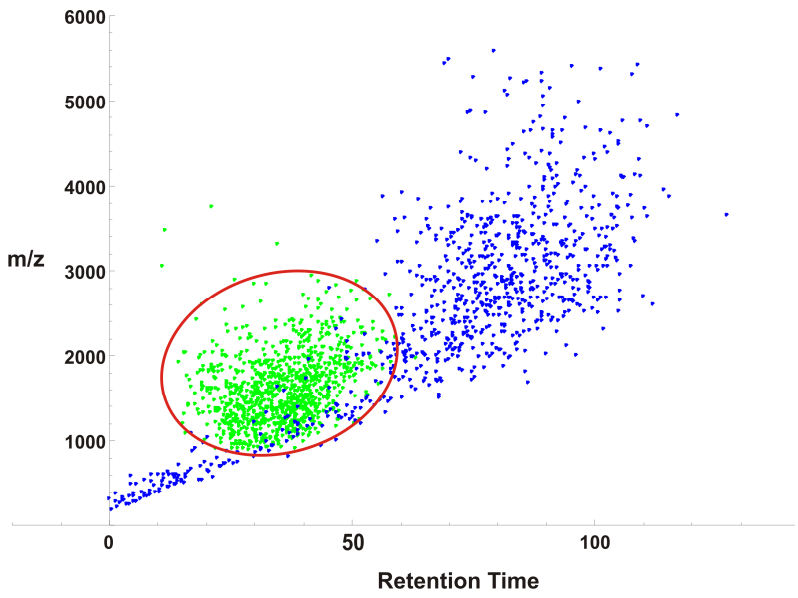


Figure 4. Peptide accessibility: transmembrane versus soluble peptides
A theoretical footprint of the peptides from the hydrophobic, transmembrane stretches of the *A. thaliana* membrane proteins (blue) is superimposed on top of the experimentally observed soluble and membrane peptides (green). The peptide masses (as m/z values) are plotted versus their normalized theoretical retention time in minutes on a reversed phase LC column; the latter is roughly proportional to their hydrophobicity.[24] The red oval roughly marks the 'hotspot' within which peptides are successfully identified by LC-MALDI-TOF/TOF. The majority of the transmembrane peptides fall outside of the hotspot.

2.2 Precursor ion selection: a problem of plenty

One of the key steps in the shotgun proteomics pipeline is the selection of precursor ions, peptides ionized in the MS mode, for subsequent fragmentation analysis in the MS/MS mode. The challenge lies in the fact that due to sample depletion (in case of off-line LC-MALDI-MS) and time constraints (on-line LC-electrospray-MS), it is not possible to select all the precursors detected by the MS mode. Subsequently, selection is often stochastic in nature and inherently biased towards abundant peptides. For a given proteome, it should be theoretically possible to identify all the proteins with a certain minimum number of peptides (a minimum of two peptides per protein is usually used as threshold for protein identification in a complex sample). So the challenge becomes one

of selecting precursor ions in a manner that it maximizes the likelihood of identifying the entire proteome in a sample. There are several factors that play an important role in this, such as peptide abundance (actual concentration), signal-to-noise ratio (instrument response), accuracy of peak picking algorithms, and occurrence of overlapping peptides (mixed MS/MS spectra). This calls for data-driven selection strategies that can cope with these factors. For instance, repeating the selection step, while excluding precursor ions selected in the previous replicates based on m/z and retention time information, has been shown to increase the total number of proteins identified.[14] A drawback of these iterative methods is the requirement for multiple analyses of the same sample or sample replicates.

2.3 Database search engines: a question of reliability

Another source of protein identification data loss (or incomplete data analysis) stems from the database search engines in the form of missed identifications or false negatives, as well as false positives. The commonly used search programs, while performing quite impressively, are far from perfect and can be sensitive to various factors such as peptide modifications, contaminants (noise), and protein database quality. Left unaccounted for, unexpected peptide modifications (experimental) lead to a mismatch with the theoretical database of peptides. In the same manner, errors in a protein database (theoretical) lead to a mismatch with the experimental MS spectra. Both of these complications can cause missed or false identifications. This is evident from the fact that different search engines often lead to different protein and peptide assignments from the same experimental data.[27,28] Thus, the reliability of search engine results is a critical issue, especially when taking into account missed assignments. Improvements in search engines are needed for creating greater accuracy and sensitivity, and thus reproducibility, in mass spectrometry-based proteomics.[29]

2.4 iTRAQ quantification: a question of significance

When faced with a list of identified proteins and their corresponding iTRAQ ratios the big question is how to differentiate between biologically significant and random events. In an ideal scenario, an iTRAQ ratio above 1.0 implies enrichment, and a ratio below 1.0 implies depletion. However, systematic and random errors can make the analysis not as straight-forward. While classical p-value based analysis is often used to gain confidence in the results, it is inappropriate in a high-throughput experiment.

This is because the p-value is only statistically valid when a single score is computed. If multiple comparisons are made, it becomes more likely that an observation will meet the p-value criteria by chance. As such, in a high-throughput analysis such as for shotgun proteomics, the problem of multiple testing can lead to false-discovery of significantly expressed proteins.[30]

Furthermore, in a quantitative experiment it might be worthwhile to pursue a more in-depth analysis of the observed proteins and their expression levels, for instance to determine whether entire classes of proteins undergo a concerted significant change in abundance. This is particularly useful when analyzing data from subtractive proteomics which involves comparing an enriched sample with a crude sample to identify proteins localized to an organelle of interest over contaminants. Access to both good protein annotation and proper statistical methods plays an important role in performing such analyses.

3. Outline of the thesis

The work done in this thesis is related to the improvement in the performance of an LC-MALDI-based proteomics pipeline by focusing on several of its components (Figure 5). This has led to several algorithms and computer programs which increase the amount of valid (**chapter II**) and reliable (**chapter III**) data and aid in its interpretation (**chapter IV, chapter V**).

In chapter II, the focus is on one of the early stages of the proteomics pipeline: selection of peptide precursor ions for fragmentation analysis. It presents a new method, APECS, which exploits the two-staged MS and MS/MS approach of a MALDI-TOF/TOF instrument to make a smarter selection by excluding redundant precursor ions. Upon validation with a complex proteome sample of *Arabidopsis thaliana*, our method identified an equivalent number of peptides as a conventional data-dependent acquisition method but with a 35% smaller work-load. While a smaller work-load translates to less analysis time, it can also be used to target precursors which would have otherwise been ignored due to sample depletion.

Moving down the pipeline, chapter III revisits the traditional scoring system used for identification of peptides. While peptides are generally identified based on the m/z information of precursor and fragment ions, there is a growing consensus that intensity-based information of fragment ions should also be included in the scoring system.[31-35] We have analyzed fragmentation patterns of peptides tagged with an iTRAQ

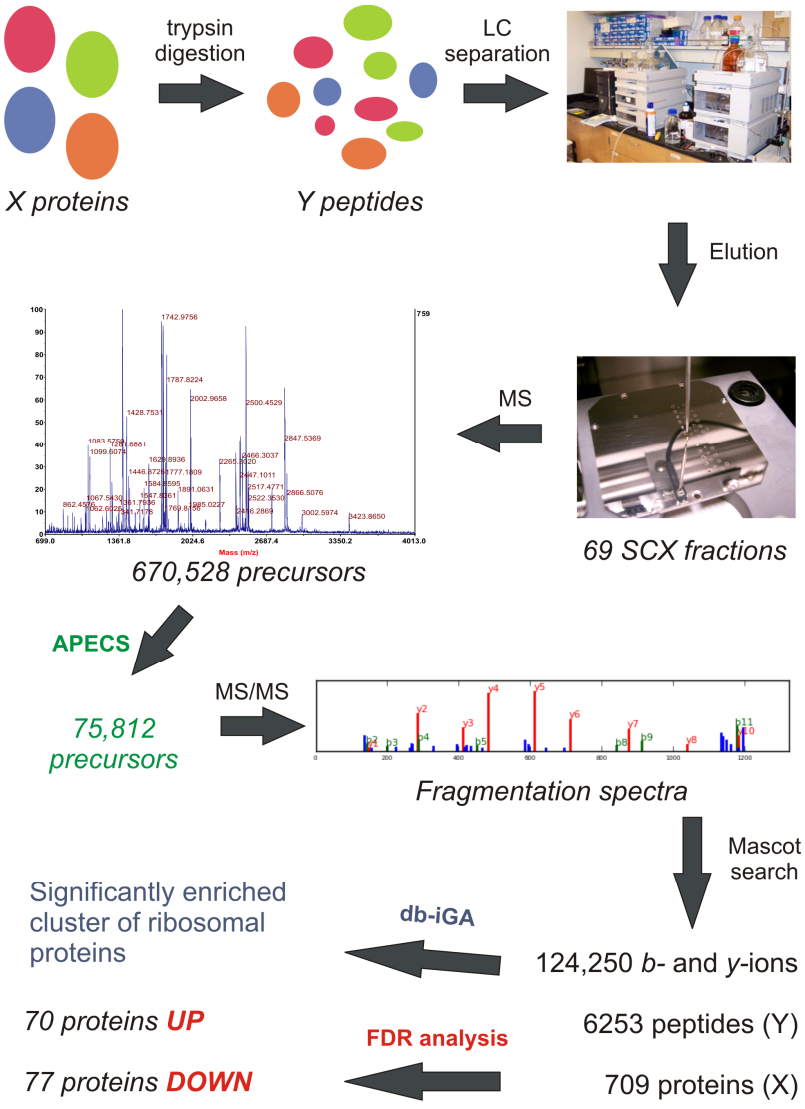


Figure 5. Scheme of an LC-MALDI-TOF-TOF experiment on the *L. lactis* proteome

Different stages in the shotgun proteomics pipeline of a quantitative analysis of the *L. lactis* membrane proteome, with typical numbers in the data acquisition stages (LC fractions, precursors, fragments) and data analysis stages (peptides, proteins, differentially regulated proteins). Apex Precursor Eluction Chain Selection (APECS), double-boundary iterative Group Analysis (db-iGA), and False Discovery Rate (FDR) analysis steps are described in chapters II, IV, and V, respectively.

label and fragmented by MALDI-TOF/TOF MS and provide further support for the idea of using the MS/MS fragment intensity for improved identification success.

Finally, chapter IV and V deal with the interpretation of quantitative results obtained with iTRAQ labeling. In chapter IV, we present the double-boundary iterative group analysis (db-iGA) method for detecting expression changes in functional classes of proteins. This is followed in chapter V by a methodology to correct for multiple testing in quantitative proteomics by performing the false discovery rate analysis (FDR).

4. References

1. Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., Hochstrasser, D. F., From proteins to proteomes: large scale protein identification by two–dimensional electrophoresis and arnino acid analysis. ***Nat. Biotech.*** 1996, 14, (1), 61–65.
2. Anderson, N. L., Anderson, N. G., Proteome and proteomics: new technologies, new concepts, and new words. ***Electrophoresis*** 1998, 19, (11), 1853–1861.
3. Washburn, M., Wolters, D., Yates, J., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. ***Nat. Biotech.*** 2001, 19, (3), 242–247.
4. Wolters, D. A., Washburn, M. P., Yates, J. R., An automated multidimensional protein identification technology for shotgun proteomics. ***Anal. Chem.*** 2001, 73, (23), 5683–5690.
5. Chen, E., Hewel, J., Felding–Habermann, B., Yates, J., Large scale protein profiling by combination of protein fractionation and multidimensional protein identification technology (MudPIT). ***Mol. Cell. Proteomics*** 2006, 5, (1), 53–56.
6. Olsen, J. V., Ong, S.-E., Mann, M., Trypsin cleaves exclusively C-terminal to arginine and lysine residues. ***Mol. Cell. Proteomics*** 2004, 3, (6), 608–614.
7. Perkins, D. N. Probability–based protein identification by searching sequence databases using mass spectrometry data. ***Electrophoresis*** 1999, 20, 3551–3567.
8. Eng, J., McCormack, A., Yates, J., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. ***J. Am. Soc. Mass Spectrom.*** 1994, 5, (11), 976–989.
9. Zubarev, R. A., Håkansson, P., Sundqvist, B., Accurate monoisotopic mass measurements of peptides: possibilities and limitations of high resolution time–of–flight particle desorption mass spectrometry. ***Rapid Commun. Mass Spectrom.*** 1996, 10, (11), 1386–1392.
10. Johnson, R. S., Martin, S. A., Biemann, K., Stults, J. T., Watson, J. T., Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. ***Anal. Chem.*** 1987, 59, (21), 2621–2625.
11. Papayannopoulos, I. A., The interpretation of collision-induced dissociation tandem mass spectra of peptides. ***Mass Spectrom. Rev.*** 1995, 14, (1), 49–73.
12. Roepstorff, P., Fohlman, J., Proposal for a common nomenclature for sequence ions in mass spectra of peptides. ***Biomed. Mass Spectrum.*** 1984, 11, (11).
13. Mueller, L. N., Brusniak, M.-Y., Mani, D. R., Aebersold, R., An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. ***J. Proteome Res.*** 2008, 7, (1), 51–61.
14. Chen, H. S., Rejtar, T., Andreev, V., Moskovets, E., Karger, B. L. Enhanced characterization of complex proteomic samples using LC–MALDI MS/MS: Exclusion

of redundant peptides from MS/MS analysis in replicate runs. ***Anal. Chem.*** 2005, 77, 7816–7825.
15. Picotti, P., Aebersold, R., Domon, B. The implications of proteolytic background for shotgun proteomics. ***Mol. Cell. Proteomics*** 2007, 6, 1589–1598.
16. Kicman, A. T., Parkin, M. C., Iles, R. K., An introduction to mass spectrometry based proteomics-Detection and characterization of gonadotropins and related molecules. ***Mol. Cell. Proteomics*** 2007, 260–262, 212–227.
17. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. ***Anal. Bioanal. Chem.*** 2007, 389, (4), 1017–1031.
18. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., Pappin, D. J., Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. ***Mol. Cell. Proteomics*** 2004, 3, (12), 1154–1169.
19. Choe, L., D'Ascenzo, M., Relkin, N. R., Pappin, D., Ross, P., Williamson, B., Guertin, S., Pribil, P., Lee, K. H., 8-Plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. ***Proteomics*** 2007, 7, (20), 3651–3660.
20. Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E. W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P. G. A., Malmstrom, J., Koehler, K., Schrimpf, S., Krijgsveld, J., Kregenow, F., Heck, A. J. R., Hafen, E., Schlapbach, R., Aebersold, R. A high-quality catalog of the *Drosophila melanogaster* proteome. ***Nat. Biotechnol.*** 2007, 25, 576–583.
21. de Godoy, L. M. F., Olsen, J. V., de Souza, G. A., Li, G., Mortensen, P, Mann, M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. ***Genome Biol.*** 2006, 7, R50.
22. Premsler, T., Zahedi, R.P., Lewandrowski, U., Sickmann, A. Recent advances in yeast organelle and membrane proteomics. ***Proteomics*** 2009, 9, 4731–4743.
23. Wiederhold, E., Veenhoff, L. M., Poolman, B., Slotboom, D. J. Proteomics of *Saccharomyces cerevisiae* organelles. ***Mol. Cell. Proteomics*** 2010, 9, 431–445.
24. Krokhin, O. V., Craig, R., Spicer, V., Ens, W., Standing, K. G., Beavis, R. C., Wilkins, J. A., An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC. ***Mol. Cell. Proteomics*** 2004, 3, (9), 908–919.
25. Wu, C. C., MacCoss, M. J., Howell, K. E., Yates, J. R., A method for the comprehensive proteomic analysis of membrane proteins. ***Nat. Biotech.*** 2003, 21, (5), 532–538.
26. Eichacker, L., Granvogle, B., Mirus, O., Müller, B. C., Miess, C., Schlieff, E., Hiding behind hydrophobicity. ***J Biol Chem*** 2004, 279, (49).
27. Kapp, E. A., Schütz, F., Connolly, L. M., Chakel, J. A., Meza, J. E., Miller, C. A., Fenyo, D., Eng, J. K., Adkins, J. N., Omenn, G. S., Simpson, R. J., An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. ***Proteomics*** 2005, 5, (13), 3475–3490.

28. Elias, J. E., Haas, W., Faherty, B. K., Gygi, S. P., Comparative evaluation of mass spectrometry platforms used in large–scale proteomics investigations. ***Nat. Meth.*** 2005, 2, (9), 667–675.
29. Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E., Beavis, R., Sechi, S., Nilsson, T., Bergeron, J. J. M., A HUPO test sample study reveals common problems in mass spectrometry–based proteomics. ***Nat. Meth.*** 2009, 6, (6), 423–430.
30. Noble, W. S., How does multiple testing correction work? ***Nat. Biotech.*** 2009, 27, (12), 1135–1137.
31. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., Gygi, S. P., Intensity-based protein identification by machine learning from a library of tandem mass spectra. ***Nat. Biotech.*** 2004, 22, (2), 214–219.
32. Zhang, Z., Prediction of low-energy collision-induced dissociation spectra of peptides. ***Anal. Chem.*** 2004, 76, (14), 3908–3922.
33. Huang, Y., Triscari, J. M., Tseng, G. C., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., Wysocki, V. H., Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. ***Anal. Chem.*** 2005, 77, (18), 5800–5813.
34. Khatun, J., Ramkissoon, K., Giddings, M. C., Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. ***Anal. Chem.*** 2007, 79, (8), 3032–3040.
35. Barton, S. J., Richardson, S., Perkins, D. N., Bellahn, I., Bryant, T. N., Whittaker, J. C., Using statistical models to identify factors that have a role in defining the abundance of ions produced by tandem MS. ***Anal. Chem.*** 2007, 79, (15), 5601–5607.

Chapter II

APECS: A new strategy for selecting precursors in 2D-LC-MALDI-TOF/TOF experiments on complex biological samples

Published in *J. Proteome Res.* 9, 5922-5928 (2010)

Abstract

LC-MALDI provides an often overlooked opportunity to exploit the separation between LC-MS and MS/MS stages of a 2D-LC-MS-based proteomics experiment, that is, by making a smarter selection for precursor fragmentation. Apex Peptide Elution Chain Selection (APECS) is a simple and powerful method for intensity-based peptide selection in a complex sample separated by 2D-LC, using a MALDI-TOF/TOF instrument. It removes the peptide redundancy present in the adjacent first-dimension (typically strong cation-exchange, SCX) fractions by constructing peptide elution profiles that link the precursor ions of the same peptide across SCX fractions. Subsequently, the precursor ion most likely to fragment successfully in a given profile is selected for fragmentation analysis, selecting on precursor intensity and absence of adjacent ions that may co-fragment. In order to make the method independent of experiment-specific tolerance criteria, we introduce the concept of the branching factor, which measures the likelihood of false clustering of precursor ions based on past experiments. Validation with a complex proteome sample of *Arabidopsis thaliana*, APECS identified an equivalent number of peptides as a conventional data-dependent acquisition method but with a 35% smaller work-load. Consequently, reduced sample depletion allowed further selection of lower signal-to-noise ratio precursor ions, leading to a larger number of identified unique peptides.

1. Introduction

Shotgun proteomics is an indispensible tool in high-throughput analysis of proteins in complex biological samples. Accurate peptide identification from liquid chromatography (LC) coupled with tandem mass spectrometry (MS/MS) forms the cornerstone of such analyses. LC-MS/MS is widely used not only for identification but also for quantification of proteins, using either isotopically labelled peptides or label-free quantification.[1] Although the analytical set-up, such as LC coupled on-line with the mass spectrometer or off-line fractionation of peptides, might differ between various shotgun proteomics strategies, there are several common steps.[2] Initially, the protein sample is treated with a protease, typically trypsin, to obtain a complex mixture of peptides. The peptides are then separated using a one- or multi-dimensional fractionation, typically chromatography, and each fraction is analyzed by MS and MS/MS. The collected MS/MS spectra are used to identify proteins, using database search programs such as Mascot.[3]

Despite the successful identification of increasingly large numbers of proteins using proteomics strategies, methodology-related constraints and the underlying complexity of a typical proteome prevent comprehensive proteome coverage.[4, 5] While advances in LC and MS instruments have improved the situation, analysis of very complex proteomes still remains a challenge. For instance, even the relatively small yeast proteome has not been fully covered.[6, 7] A major bottleneck remains the problem of plenty; a typical two-dimensional (2D) LC-MS experiment of a complex sample will result in hundreds of thousands of peaks with a good signal-to-noise ratio. This makes it unfeasible to perform MS/MS fragmentation on all of them due to constraints in time and sample amount. As a result, considerable effort has been directed towards improving the work flow of such experiments to boost the proteome coverage. The established methods for precursor ion selection, referred to as data-dependent acquisition (DDA), work by selecting the most intense peptide signals from each MS spectrum for MS/MS analysis. As such, these methods fail to account for the redundancy resulting from abundant peptides. It has been previously suggested that using a two-staged approach, where the sample is first analyzed using LC-MS and then the identification is performed using an inclusion list for MS/MS based on the first step, improves the coverage.[8] This makes matrix-assisted laser desorption ionization (MALDI) mass spectrometry, which already has decoupled these two stages, especially suitable. Yet, this inherent advantage offered by MALDI is often overlooked by using a DDA selection strategy.

To address the problem of redundancy, various strategies have been reported for selecting precursor ions, all of which rely on giving priority to precursor ions with a high potential of revealing new information. Most of these methods aim to resolve redundant data acquisition by minimizing fragmentation of multiple peptides from abundant proteins. Mass-based dynamic exclusion list strategies which exclude peptides that are already fragmented or stem from identified proteins have been shown to improve the number of unique identifications for both MALDI and electrospray ionization (ESI).[9-11] Another approach, the so-called directed mass spectrometry, makes use of prior information to build a compound specific profile of expected peptides. This profile then forms the basis for precursor ion selection by building an inclusion list for a nonredundant analysis.[12-14] The directed approach was combined with an iterative strategy where precursor ions are assigned a dynamic weighting factor based upon the uniqueness of a precursor ion mass in a particular proteome.[15]

Another source of redundancy arises from the fact that a precursor ion might elute in a broad LC peak, over multiple LC-MALDI fractions. This often results in identical precursor ions being selected for MS/MS fragmentation, while other, concurrently eluting, precursor ions are not analysed. Exclusion algorithms are employed to resolve this issue, but they rely on arbitrary criteria such as user-defined number of times a peak should be excluded from fragmentation analysis. A chromatographic peak model to resolve the redundancy arising from abundant precursors with strong peak tailing has been shown to work for on-line LC-ESI MS analyses, where precursor ion selection has to be done on the fly.[16] Although peptide exclusion from adjacent fractions in a single LC-MALDI run is possible with existing data acquisition software, this has not been extended to multi-dimensional separations.

Here, we present Apex Peptide Elution Chain Selection (APECS), a simple and powerful method for intensity-based peptide selection in a complex sample separated by 2D-LC using a MALDI-TOF/TOF instrument. APECS aims to exploit the decoupling advantage offered by a MALDI instrument to remove the peptide redundancy present in the adjacent fractions of the first LC dimension. This is achieved by constructing elution profiles of all peptides across all fractions. This information is useful in itself for precursor ion selection, as shown here, but can also be used in a complementary manner with the various strategies mentioned above. APECS was experimentally tested on a sample from the *Arabidopsis thaliana* proteome using a 2D-LC MALDI-TOF/TOF setup. Comparison with DDA selection strategies shows APECS requiring a 35% smaller work-load with a small gain in the total number of distinct peptides identified.

2. Material and Methods

2.1 Sample preparation and 2D-LC-MS/MS

Two different datasets were used for the evaluation of different approaches. Dataset 1 was taken from a published work involving *Lactococcus lactis* membrane samples.[17] Dataset 2 consisted of two independent membrane samples extracted from NaCl-stressed *A. thaliana* plants labelled with 4-plex iTRAQ and analyzed as follows. After acetone precipitation each sample (100 μg) was resuspended in 40 □L of 500 mM triethylammonium bicarbonate (TEAB), 0.05% sodium dodecylsulfate (SDS). Cysteine modification with methyl-methanethiosulfonate (MMTS), digestion with trypsin (Cat. V511A, Promega) and 4-plex iTRAQ-labelling were performed according to the

manufacturer's instructions (Applied Biosystems, Foster City, CA, USA). After labelling the samples were pooled with equal protein ratio and lyophilized.

A silica-based Polysulfoethyl Aspartamide strong cation exchange (SCX) column (Cat. 202SE0502, PolyLC Inc., Columbia USA) was used for off-line peptide separation on an Ettan-MDLC system (Amersham Biosciences AB, Uppsala, Sweden) at a flow rate of 200 µL/min with UV detection. Buffer A contained 10 mM KH_2PO_4-H_3PO_4, pH 2.7, 25% acetonitrile (ACN) and buffer B contained 10 mM KH_2PO_4-H_3PO_4, pH 2.7, 25% ACN, 1 M KCl. Pooled iTRAQ labelled samples were resuspended in buffer A prior to loading. Peptide elution was performed with a step gradient from 3 to 12% B in 12 CV (column volumes), followed by 12 to 30% B in 3 CV. Fractions were collected every 45 sec in 96-well plates. Eluted peptides were first vacuum dried to remove the ACN excess and subsequently diluted with 0.1% trifluoroacetic acid (TFA). Depending on the complexity as judged by UV signal intensity, either separate fractions, or pools of two fractions, were analyzed by reverse phase LC-MALDI-TOF/TOF.[17] Fractions of 12 seconds were spotted on a blank MALDI target with a Probot system (LC Packings, Amsterdam, The Netherlands). Mass Spectrometric analysis was carried out with a 4800 Proteomics Analyzer MALDI-TOF/TOF instrument (Applied Biosystems) in the m/z range 900-5000. Data acquisition was performed in positive ion mode. Peptides were selected for MS/MS fragmentation using the APECS method described below. Protein identifications were confirmed using Mascot (Matrix Science, London, UK; version 2.1) and the TAIR7 sequence database.[18] All peptide matches with a confidence of identification higher than 95% were accepted.

2.2 Apex peptide elution chain selection

The aim of APECS is that, for a peptide (precursor) that elutes over multiple first-dimension LC fractions, only the chromatographic peak fraction will be sampled for MS/MS fragmentation. The LC method for first-dimension separation is typically strong cation exchange (SCX), but APECS works for any LC method. APECS links precursor ions together in a chain and then selects the one with the highest signal-to-noise ratio (SNR), referred to as Apex Precursor Ion. Tolerance of mass and second-dimension LC retention time determine the linking of peptides across SCX fractions. For large tolerances multiple candidate precursors may emerge for a specific peptide already present in the previous run (Figure 1). However, only one candidate is assumed

to be correct. Due to the low sampling rate in the SCX dimension (45 s or 1 min in our datasets), discrimination of candidate precursors based on the intensity profile is not possible. APECS therefore builds tree-like clusters consisting of all the possible matching candidates in each SCX run. We call these Peptide Elution Chains and Peptide Elution Trees, which together form the Peptide Elution Profile.

A Peptide Elution Chain (chain) defines the elution profile of a specific peptide by connecting precursors within the tolerance from subsequent SCX fractions. A chain may contain one or more precursors, and no SCX fraction gaps are allowed between any two subsequent precursors. A Peptide Elution Tree (tree) defines a cluster of one or more peptide elution chains, with one start node (Figure 1). In its simplest form, a tree represents a single-precursor chain. In more complex forms, it may contain multiple chains with multiple branching points representing a mixed elution profile.

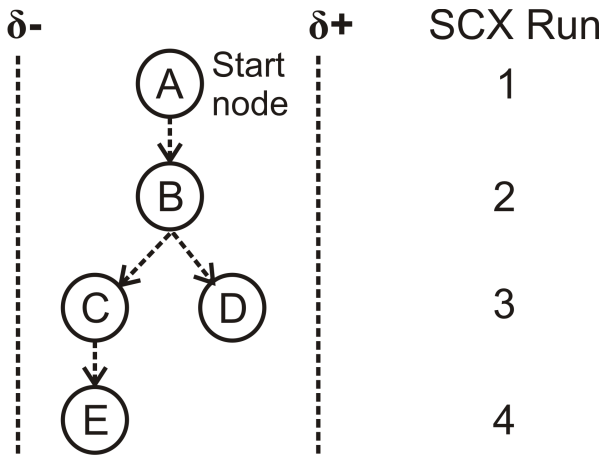


Figure 1. A tree schematic representing the elution profiles of two different peptides, in subsequent first-dimension SCX runs.

Each node represents a precursor ion found in that run, within the defined mass and second-dimension LC retention time tolerances (δ- and δ+). Precursor ion A represents the start node of the tree, whereas precursor ions D and E represent the end nodes. Precursor ions A and B form an ambiguous path as they could be part of a peptide elution profile represented by A→B→D or A→B→C→E.

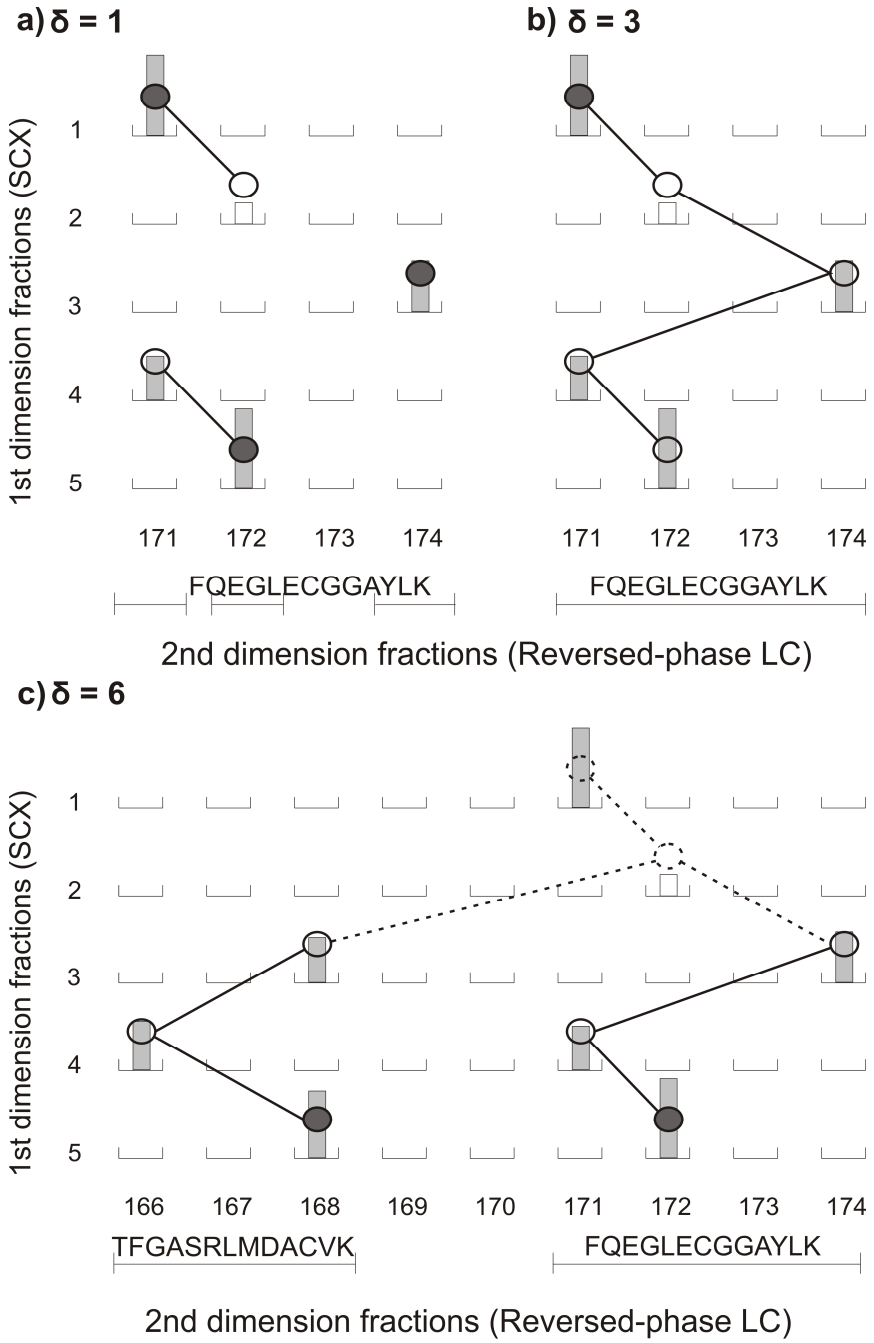


Figure 2. Schematic example of the effect of second-dimension retention time tolerance on peptide elution profiles

The effect of increasing the second-dimension retention time tolerance δ in constructing the elution profile of peptide FQEGLECGGAYLK from *A. thaliana*: a) 12 s tolerance (1 fraction), b) 36 s (3 fractions), c) 72 s (6 fractions). The height of the bars represents the intensity (as signal-to-noise ratio) of the precursor, and the gray bars are ranked within the top ten most intense peaks in that fraction. The connected circular nodes represent the elution trees; a solid black node indicates the Apex Precursor Ion, a dashed connection indicates an ambiguous path. A δ of 1 fraction leads to formation of 3 isolated elution trees, and therefore 3 Apex Precursor Ions (a). At a δ of 3 fractions all of the precursor ions form a single chain, leading to a single Apex Precursor Ion (b). A δ of 6 fractions causes formation of a branched tree linked with precursors identified as a different peptide, TFGASRLMDACVK (c). The Apex Precursor Ions are selected only from the unambiguous path, so both peptides are still identified, although for FQEGLECGGAYLK the SNR in SCX fraction 1 is slightly higher than in fraction 5.

From a Peptide Elution Profile the Apex Precursor Ion is selected for MS/MS fragmentation. In chains, this is simply the fraction with the highest SNR. In trees with multiple branches, the precursors with the highest SNR from the unshared path(s) (i.e. after the last branching point) are selected. A higher number of complex trees form as the mass and retention time tolerances are increased (Figure 2). A complex tree therefore represents two or more peptides whose elution profiles cannot be discriminated by the chosen tolerance criteria. The branching factor, i.e. the ratio of the number of chains over the number of trees, can be used to qualify the discriminating power of the chosen tolerance criteria, since it gives a measure of how often a certain tolerance leads to a false clustering of distinct precursors. In complex datasets, it is expected that the rate of branching is directly proportional to the rate of false clustering. Calculation of the branching factor allows for a more intuitive assessment of the false clustering rate than dealing directly with tolerance parameters.

2.3 Implementation details

Preprocessing

After 2D-LC separation of a sample, MS analysis of all runs is performed. The instrument acquisition software of the 4800 Proteomics Analyzer uses an Interpretation Method to determine the apex fraction of each precursor present in adjacent spots for each individual (second-dimension) LC-MALDI-TOF run. In case such a method is not

available for a particular MALDI-TOF/TOF instrument, APECS can easily be adapted to determine apex fractions in the second dimension as well. The filtered lists of candidate precursors from each (second-dimension) LC-MALDI-TOF run are then preprocessed to flag all precursors overlapping in individual spectra within a mass resolution window of 300 for the m/z range of 2000 to 4000 and 200 for the m/z range of 900 to 2000. Overlapping precursors often result in undesirable, mixed fragmentation spectra and will be excluded for MS/MS analysis, but only after elution profile construction.[19]

Building trees

New trees are created for each precursor ion in the first second-dimension LC run of the first SCX fraction. These precursor ions act as the root nodes of their respective trees. Subsequently, each precursor ion from the next LC run is compared with the average mass and retention time of active chains from the previous LC run within the tolerance window. If a precursor ion matches an existing chain, then it is linked to it, thereby extending it. Conversely, if a precursor ion is not matched to any of the existing active chains, then it becomes the root node of a new tree. If more than one precursor ion from the current LC run matches the same active chain, then a branched path is created.

SCX fraction gaps between two subsequent precursors are not allowed in a peptide elution chain in a tree. Therefore, all chains that have not been extended at the end of an extension step are pruned out. Each such chain is traced back until either a parent precursor ion with two or more child precursor ions (i.e. a branching point) is encountered or the root node is reached. The link is then severed between this parent precursor ion and the chain. In the case of reaching the root node, the entire tree is effectively pruned out.

Selecting Apex Precursor Ions

After a chain is pruned out of a tree, the precursor ion with the highest SNR is selected for fragmentation given that it fulfills two conditions. The precursor ion should not be labeled as overlapping during the preprocessing step and not have an ambiguous chain membership, unless it is the only candidate with a SNR above the pre-specified minimum. The branching factor is then calculated as a quality check for the tolerance parameters, by dividing the total number of trees by the total number of branches from all the trees. Using the same tolerance criteria for two samples of different complexity

would lead to a lower branching factor for the sample with the higher complexity. In such cases, reprocessing can be done with more conservative tolerance criteria for the more complex sample. Finally, a list of selected Apex Precursor Ions, one per elution chain, is compiled which forms an inclusion list for the MS/MS fragmentation.

3. Results and discussion

3.1 *In silico* analysis of an experiment ran without APECS

Data acquisition of a 2D-LC-MS proteome

A *L. lactis* membrane proteome dataset[17] contained, across 69 SCX fractions, 131,430 candidate precursor ions with a SNR above the pre-specified threshold of 120 for peptides within the m/z range of 900 to 2000 and of 50 for peptides within the m/z range of 2000 to 4000. A brute-force approach, where all of these candidates are fragmented (DDA-ALL), is usually not feasible or productive. This is due to time constraints and loss of material by scanning the same spot multiple times, resulting in progressive degradation of the quality of subsequent spectra. In practice, an upper limit to the number of precursors per spot, usually ten, is imposed (DDA-TOP10). In a different approach to reduce time and depletion constraints, a mass-dependent selection strategy (DDA-ALT) is routinely employed in our group, alternating between a high and a low mass range.[17, 20] For each odd second-dimension (reversed phase) LC run the 15 most intense peaks per spot above the SNR of 120 are selected for MS/MS fragmentation in the m/z range from 900 to 2000, whereas in the even LC runs the 10 most intense peaks above the SNR of 50 in the m/z range from 2000 to 4000 are selected. The DDA-ALT strategy in effect skips every other SCX fraction, but at the expense of undersampling, depending on the actual SCX peak widths of individual peptides. With this selection strategy, 85,677 precursor ions (65% of the total pool) were discarded from fragmentation analysis (Table 1). The remaining 45,753 precursor ions were then measured in approximately 101 hours of MS/MS analysis time.

Selecting the tolerance parameters for APECS

The *L. lactis* membrane proteome dataset was reanalyzed to calculate the branching factor with four different retention time and m/z tolerance criteria (Figure 3). In addition, the corresponding False Clustering Rate (FCR) and Work Load (WL) were

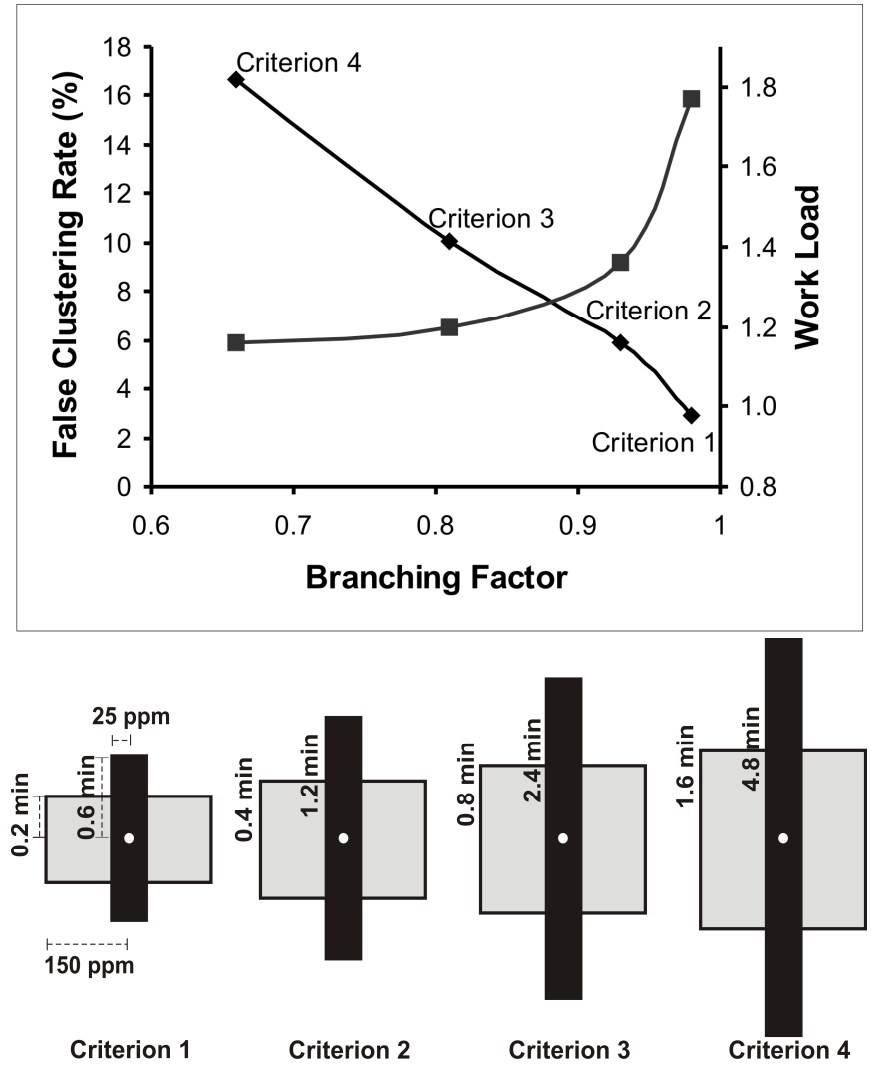


Figure 3. Relationship between branching factor, false clustering rate and work load in terms of mass (m/z) and retention time tolerance criteria

Effect of different tolerance criteria for constructing elution trees on the False Clustering Rate (left y-axis, diamonds), and the Work Load (right y-axis, squares). Criteria 1 to 4 lead to Branching Factors of 0.98, 0.93, 0.81, and 0.66, respectively, for the *L. lactis* dataset. For each criterion a combination of a small tolerance for the second-dimension LC retention time (Rt1) at a mass tolerance of 150 ppm, and a large retention time tolerance (Rt2, not drawn to scale) at a small mass tolerance of 25 ppm,

are used, to account for fluctuations in mass calibration and LC stability. Criterion 1: Rt1 <= 0.2 min, Rt2 <= 0.6 min; criterion 2: Rt1 <= 0.4 min, Rt2 <= 1.2 min; criterion 3: Rt1 <= 0.8 min, Rt2 <= 2.4 min; criterion 4: Rt1 <= 1.6 min, Rt2 <= 4.8 min.

calculated. The FCR corresponds to the rate of falsely-linked precursor ions belonging to different peptides, as identified by Mascot. FCR is calculated by dividing the total number of precursors that were clustered to the same peptide but were identified as different peptides by the total number of different identified peptides. The Work Load is the number of precursors selected for fragmentation divided by the number of redundant, discarded precursors.

As shown in Figure 3, the branching factor is inversely proportional to the False Clustering Rate. This is not surprising since a lower branching factor implies a more liberal choice of tolerance parameters. At the same time, the branching factor is directly proportional to the number of precursors that are selected for fragmentation, and by extension, to the Work Load. The branching factor of 0.93 at criterion 2 represents the point after which the Work Load increases rapidly in respect to the improvement in the FCR. The branching factor therefore provides an objective criterion for controlling the quality of the Peptide Elution Chains formed using APECS.

Elution Profile Analysis

Construction and analysis of the elution profiles of all 131,430 precursor ions in the SCX dimension with APECS using criterion 2 revealed that 75,812 (58%) were assigned as Apex Precursor Ions (Table 1). Assuming these to be the complete set of unique peptides, 8,375 (18.3%) of the 45,753 precursor ions selected using DDA-ALT were redundant. In terms of time, this represents an estimated 18 hours of MS/MS analysis which could have been spent more efficiently by analyzing more of the discarded candidates. Even more significantly, only 62% (37,378) of the Apex Precursor Ions are selected for fragmentation using the alternating strategy. Thus, while the latter strategy also reduces the acquisition time, it does so at a significant cost to the diversity of the selection pool. Elution profiles on the other hand provide a context to the precursor ions in different SCX fractions, making it possible to select for fragmentation in a systematic manner.

3.2 Experimental validation of the elution profile strategy

Peptide Elution Profiles were created for the *A. thaliana* sample using the tolerance parameters from criterion 2 (Figure 3), resulting in 217,934 chains and 202,693 trees. The calculated branching factor of 0.93 is the same as that obtained from the *L. lactis* dataset with the same criterion. After applying the same SNR filter as for the *L. lactis* dataset 39,603 precursor ions were selected for fragmentation based on elution profiles. An additional 39,743 precursor ions were selected from the top 10 most intense peaks per fraction, which were found to be redundant. In this manner, a comparison can be made within the same dataset between the elution profile selection strategy using APECS, and the intensity-based selection using the conventional DDA-TOP10 strategy. A total of 79,346 precursor ions were submitted for fragmentation, and subsequent Mascot-based identification found 20,081 peptides with a confidence of greater than 95%.

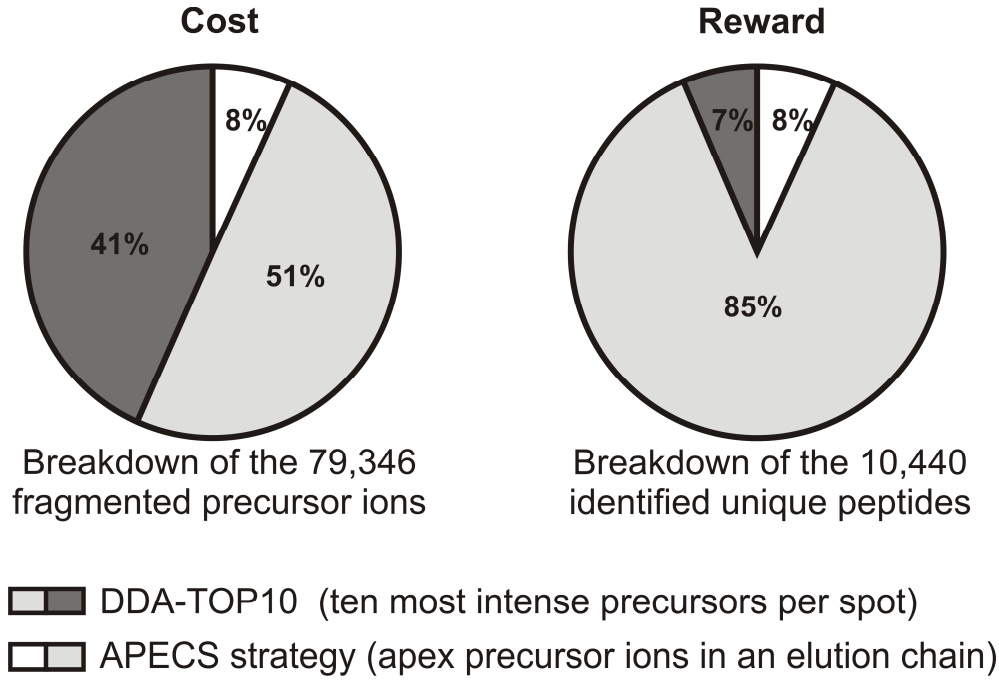


Figure 4. A cost versus reward comparison of DDA-TOP10 and APECS strategy
Breakdown of number of peptides acquired (left diagram, cost) and identified (right diagram, reward) using either the APECS elution profile strategy for precursor ion selection, or the top ten intensity-ranked selection strategy (DDA-TOP10).

A comparison of the precursor ions selected for fragmentation by the two strategies (cost) versus the number of unique peptides identified (reward) is shown in Figure 4. While the number of unique peptides identified by both methods remains similar (9,604 vs 9,709 for the DDA-TOP10 and APECS strategy, respectively), there is a large difference in the amount of work required by each, with APECS using 35% fewer precursor ions. The additional 32,531 precursors acquired by the DDA-TOP10 strategy only result in 693 unique peptides missed by APECS. The additional 800 unique peptides from APECS were relatively low SNR precursors outside of the ten most intense precursors in their respective fractions. As expected, their identification rate (13%) is markedly lower than for the more intense precursors (21%).

3.3 Advantages and limitations of APECS

The advantages of the APECS strategy are illustrated by examples of two Peptide Elution Profiles (Figure 5). Peptide SGGVTDDSGSTK elutes as a very broad peak in the SCX separation: its elution profile consists of ten precursor ions in subsequent LC runs (Figure 5a), each with a sufficiently high SNR to get selected for fragmentation by DDA-TOP10 and identified by Mascot as the same peptide. Using APECS, however, only P6 would have been selected and the other nine precursors would have been discarded as redundant without any loss of information. Figure 5b shows an example of a branching elution profile having five precursor ions in two different chains. The peptides are close enough in mass and retention time for the root node to be linked to the incorrect chain. Nevertheless, since APECS always picks a unique precursor from each branch, in this case P2.2 and P3, it was possible to discriminate both partially co-eluting peptides (LVGLVNDEETDSGR, 1502.7213 Da, and LWTNPDEFNPDR, 1502.6790 Da). The DDA-TOP10 strategy selected four precursor ions for fragmentation, P1, P2.1, P3 and P4, again without the gain of additional information. In this case, the fragmentation of P2.2 resulted in a confident identification, although it has a relatively low SNR of 56. In other cases, where multiple peptides elute within the tolerance criteria, the discriminating power of the method may be insufficient. Hence, the branching factor analysis is performed to limit the number of branching trees.

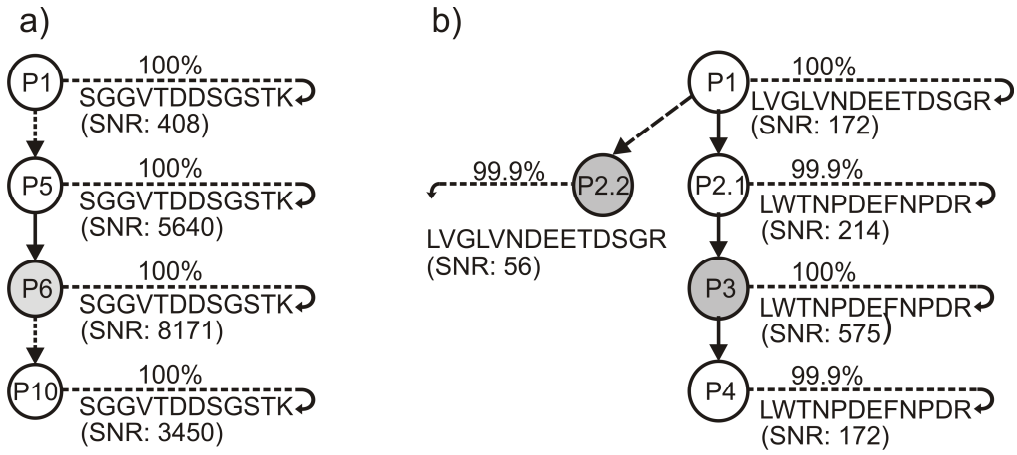


Figure 5. A schematic example of two different Peptide Elution Profiles

a) A schematic of the elution profile of a peptide from *A. thaliana* consisting of ten precursor ions (P1-P10), detected at the indicated SNR levels in ten subsequent SCX fractions. All were identified with at least 95% confidence by Mascot as SGGVTDDSGSTK. Due to the high SNR all ten precursors are selected for fragmentation by the DDA-TOP10 method. However, with the APECS elution profile strategy, only precursor ion P6 was selected and the rest discarded as redundant. b) A schematic of the elution profile of two different chains with five precursor ions (P1-P4) in four subsequent SCX fractions. In fraction 2, two potential precursors P2.1 and P2.2 can be linked to P1. After Mascot identification it became evident that P1 and P2.2 represent the same peptide, and P2.2-P4 a different one. Precursor ions P2.2 and P3, selected by APECS, suffice to identify both peptides. The DDA-TOP10 method selected four precursor ions (P1, P2.1, P3, and P4) and resulted in the same two identified peptides. P2.2 was not selected because it was not among the ten most intense peaks in its spot.

Ideally, APECS would identify all the unique peptides present in the sample. However, in the validation experiment 7% (693) of the identified unique peptides stemmed from precursor ions which were rejected by APECS as redundant. A closer look revealed that these peptides are either precursor ions wrongly clustered by APECS as part of a chain, or precursor ions with a higher quality spectrum than corresponding Apex Precursor Ion in the chain. The former is a consequence of the trade-off in False Clustering Rate and Work Load. In regards to the latter, the precursor ions are correctly

rejected as redundant, but our assumption that the highest SNR precursor ion is the best candidate for fragmentation sometimes fails, in particular where the SNR is close to the threshold. In general, however, the SNR correlates very well with identification rate (Figure 6). Another explanation is related to the rank of the precursor ion in the acquisition queue. As evident from our results, the identification rate of the precursor ions that were selected by APECS amongst the ten most intense peaks in a spot is significantly higher than of the ones outside of the top ten. Since acquisition proceeds in order of decreasing intensity, depletion of a spot combined with the lower SNR is expected to lead to lower identification success. This also means that, if not for the acquisition of redundant precursor ions in this dataset for comparison purposes, more unique peptides could have been identified from the latter group of precursor ions. The validation dataset analysis therefore underestimates the effectiveness of the APECS method.

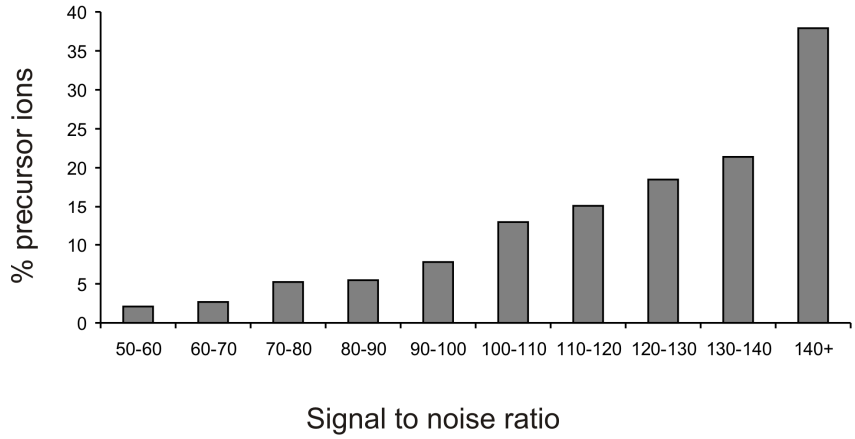


Figure 6. The effect of signal-to-noise ratio on the identification rate (Mascot confidence of > 95%) of 79,346 precursor ions from the *A. thaliana* dataset.

3.4 Scope of APECS

Creation of Peptide Elution Profiles is an effective way to perform pre-MS/MS filtering of redundant peptides. The method gains power for complex biological samples where two-dimensional fractionations are routinely performed. Even when a crude sampling rate of ten or fewer fractions per first dimensional (SCX) separation is used, overlap in fractions is hard to avoid. The higher the first-dimension sampling rate, the

more peptides may be identified,[21] but the more important filtering strategies such as APECS become to reduce the work load.

Once the elution profiles are created for each peptide in a sample, the selection criteria for the fragmentation analysis can be quite flexible. While only the Apex Precursor Ions were selected in this work, it is possible to adjust the selection process according to the context of the elution profile. For example, additional precursor ions could be selected from trees with multiple branching to compensate for potential false clustering of peptides. APECS can also be used in conjunction with other dynamic precursor ion selection strategies such as exclusion of precursor ions stemming from already identified peptides over several replicates. Currently, instrument software performs similar analyses for single LC-MALDI runs and this work makes a strong case for extending it over multi-dimensional LC-MALDI runs to account for the strong redundancy in those dimensions.

4. Conclusion

MALDI provides an opportunity to exploit the separation between LC-MS and MS/MS stages of a 2D-LC MS-based proteomics experiment by making a smarter selection for precursor fragmentation. However, this is often squandered and the physical separation of LC and MS itself and the associated lack of automation are considered drawbacks of this workflow. We have shown that having access to the 2D-LC elution profile of a peptide affords a level of flexibility that could be harnessed for a more complete coverage of a proteome. APECS identified equivalent number of peptides as with the data-dependent approach but with a 35% smaller work-load. Consequently, reduced sample depletion allows further selection of lower signal-to-noise ratio precursor ions, leading to a larger number of identified unique peptides. Full use of the advantages of the LC-MALDI approach for 2D-LC experiments of complex samples should allow a significant improvement in proteome coverage. The application software package is available for download at https://trac.nbic.nl/apecs/.

5. Acknowledgement

We thank W. Huibers for technical assistance with LC-MALDI-MS. This research work was supported by the Netherlands Bioinformatics Centre (NBIC) and the Netherlands Proteomics Centre (NPC).

6. References

1. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. ***Anal. Bioanal. Chem.*** 2007, 389*,* (4), 1017–1031.
2. Aebersold, R., Mann, M. Mass spectrometry-based proteomics. ***Nature*** 2003, 422*,* (6928), 198–207.
3. Perkins, D. N. Probability-based protein identification by searching sequence databases using mass spectrometry data. ***Electrophoresis*** 1999, 20*,* (18), 3551–3567.
4. Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E. W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P. G. A., Malmstrom, J., Koehler, K., Schrimpf, S., Krijgsveld, J., Kregenow, F., Heck, A. J. R., Hafen, E., Schlapbach, R., Aebersold, R. A high-quality catalog of the *Drosophila melanogaster* proteome. ***Nat. Biotechnol.*** 2007, 25*,* (5), 576–583.
5. de Godoy, L. M. F., Olsen, J. V., de Souza, G. A., Li, G., Mortensen, P, Mann, M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. ***Genome Biol.*** 2006*,* 7,(6), R50.
6. Premsler, T., Zahedi, R.P., Lewandrowski, U., Sickmann, A. Recent advances in yeast organelle and membrane proteomics. ***Proteomics*** 2009, 9, (20), 4731–4743.
7. Wiederhold, E., Veenhoff, L. M., Poolman, B., Slotboom, D. J. Proteomics of *Saccharomyces cerevisiae* organelles. ***Mol. Cell. Proteomics*** 2010, *9,* 431–445.
8. Picotti, P., Aebersold, R., Domon, B. The implications of proteolytic background for shotgun proteomics. ***Mol. Cell. Proteomics*** 2007, 6, (9), 1589–1598.
9. Scherl, A., Francois P., Converset, V., Bento, M., Burgess, J. A., Sanchez, J. C., Hochstrasser, D. F., Schrenzel, J., Corthals, G. L. Nonredundant mass spectrometry: a strategy to integrate mass spectrometry acquisition and analysis. ***Proteomics*** 2004, 4, (4), 917–927.
10. Chen, H. S., Rejtar, T., Andreev, V., Moskovets, E., Karger, B. L. Enhanced characterization of complex proteomic samples using LC-MALDI MS/MS: exclusion of redundant peptides from MS/MS analysis in replicate runs. ***Anal. Chem***. 2005, 77, (23), 7816–7825.
11. Wang, N., Li, L. Exploring the precursor ion exclusion feature of liquid chromatography-electrospray ionization quadrupole time-of-flight mass spectrometry for improving protein identification in shotgun proteome analysis. ***Anal. Chem.*** 2008, 80, (12), 4696–4710.
12. Rinner, O., Mueller, L. N., Hubálek, M., Müller, M., Gstaiger, M., Aebersold, R. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. ***Nat. Biotechnol.*** 2007, 25, (3), 345–352.
13. Schmidt, A., Gehlenborg, N., Bodenmiller, B., Mueller, L. N., Campbell, D., Mueller, M., Aebersold, R., Domon, B. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. ***Mol. Cell. Proteomics*** 2008, *7,* (11), 2138–2150.

14. Schmidt, A., Claassen, M., Aebersold, R. Directed mass spectrometry: towards hypothesis-driven proteomics. ***Curr. Opin. Chem. Biol.*** 2009, 13 (5-6), 510–517.
15. Zerck, A., Nordhoff, E., Resemann, A., Mirgorodskaya, E., Suckau, D., Reinert, K., Lehrach, H., Gobom, J. An iterative strategy for precursor ion selection for LC-MS/MS based shotgun proteomics. ***J. Proteome Res.*** 2009, 8, (7), 3239–3251.
16. Kohli, B. M., Eng, J. K., Nitsch, R. M., Konietzko, U. An alternative sampling algorithm for use in liquid chromatography/tandem mass spectrometry experiments. ***Rapid Commun. Mass Spectrom.*** 2005, 19, (5), 589–596.
17. Steen, A., Wiederhold, E., Gandhi, T., Breitling, R., Slotboom, D. J. Physiological adaptation of the bacterium *Lactococcus lactis* in response to the production of human CFTR. *Mol. Cell.* ***Proteomics*** 2010, doi:10.1074/mcp.M000052–MCP201.
18. Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., Zhang, P. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. ***Nucleic Acids Res.*** 2003, 31 (1), 224–228.
19. Ow, S. Y., Salim, M., Noirel, J., Evans, C., Rehman, I, Wright, P. C. iTRAQ underestimation in sample and complex mixtures: the good, the bad and the ugly. ***J. Proteome Res.*** 2002, 8, (11), 5347–55.
20. Wiederhold, E., Gandhi, T., Permentier, H. P., Breitling, R., Poolman, B., Slotboom, D. J. The yeast vacuolar membrane proteome. ***Mol. Cell. Proteomics*** 2009, 8, (2), 380–392.
21. Sandra, K., Moshir, M., D'Hondt, F., Tuytten, R., Verleysen, K., Kas, K., François, I., Sandra, P. Highly efficient peptide separations in proteomics: Part 2: Bi- and multidimensional liquid–based separation techniques. ***J. Chromatogr. B.*** 2009, 877, (11-12),1019–103

Chapter III

The effect of iTRAQ labeling on relative abundance of ions produced by tandem MALDI-MS

Abstract

While database search programs are quite efficient in identifying proteins from mass (m/z) information derived from tandem MS data, they can suffer from issues related to missed identifications and reliability. In the past, using the information available in the relative ion abundance or fragment peak intensity has been shown to have the potential to improve their performance. In this study, we consider the effect of iTRAQ labeling on the peptide fragmentation of singly-charged fragments from tandem MALDI MS. The presence of an iTRAQ modified basic group on the N-terminus leads to a more pronounced set of *b*-ion peaks. We performed a rather simple intensity prediction by using a decision-tree machine learning approach and were able to show that relative ion abundance in a spectrum can be correctly predicted (Match dataset) and distinguished from closely related sequences (Mismatch dataset).

1. Introduction

The characterization of proteins in complex biological mixtures remains a major objective in proteomics-based research. Database search driven protein identification from tandem mass spectra of peptides is the most widely used method for such analyses. Often, the spectra are derived from peptide fragmentation through a low-energy collision-induced dissociation (CID) process. In addition to identification, quantification is an important factor in monitoring changes in a proteome under different physiologically relevant conditions. With the iTRAQ (isobaric tags for absolute and relative quantification) labelling strategy, peptide fragmentation allows to concurrently perform protein identification and quantification.

Despite the successful characterization of increasingly large numbers of proteins using proteomics strategies, methodology-related constraints and the underlying complexity of a typical proteome prevent comprehensive proteome coverage.[1,2] While advances in LC and MS instruments have improved the situation, analysis of very complex proteomes still poses a hefty challenge. One of the major challenges stems from the database search engines in the form of missed identifications or false negatives. In a typical LC-MS/MS experiment of a complex proteome sample, only a small percentage of the peptide fragmentation spectra are successfully identified.[3,4] The commonly used search programs, while impressive, are far from perfect and can be sensitive to various factors such as insufficient database quality, unexpected peptide modifications, contaminants, and low spectral quality. This is

evident from the fact that different search engines often lead to different protein and peptide assignments.[5,6] Thus, performance of search engines is a critical issue to the overall performance of the strategy. In the past, an improvement in the search engine performance has been called for in order to create greater reproducibility in mass spectrometry-based proteomics.[7]

While peptides are generally identified based on their mass information (derived from the measured mass-to-charge ratio, m/z) by the various database search engines, there is a growing consensus that information about relative ion abundance (fragment intensity) should also be a criterion.[8-14] Much work has already been done to elucidate the fragmentation pathways of protonated peptides. The most significant outcome of this work has been the so-called *mobile proton model* which describes how protonated peptides dissociate upon excitation by CID, depending on proton distribution.[15-18] Peptide fragmentation primarily follows two competing pathways, "charge-directed" and "charge-remote" process. In a charge-directed process, the increase in the internal energy of the ion upon excitation leads to the migration of the sequestered proton along the peptide backbone. The migration occurs from the initial site of protonation to energetically less favored protonation sites leading to backbone dissociation. However, if the proton is tightly sequestered at the initial site of protonation such as the strongly basic arginine (Arg), then the energy required to mobilize the proton is too large leading to a charge-remote dissociation. Whether mobile or sequestered (immobile) protons are available in a particular peptide depends on the charge and amino acid composition. Accordingly, ionized peptides are classified in three ways:

1. Mobile-proton peptides in which the number of basic residues is smaller than the peptide's charge. For example, doubly-charged tryptic peptides with only a C-terminal basic lysine (Lys) or Arg residue.
2. Immobile-proton peptides in which number of Arg residues greater or equal to the peptide's charge. For example, singly-charged tryptic peptides with at least one Arg.
3. Partially mobile-proton peptides are the ones that cannot readily be classified as mobile or immobile. For example, singly-charged tryptic peptides with a single lysine.

Mobile-proton peptides heavily favour charge-directed fragmentation due to the availability of a free charge (mobile proton), whereas immobile-proton peptides tend to favour a charge-remote pathway. Partially mobile proton peptides, on the other hand, are not as selective.

In this work, we investigate the fragmentation patterns of peptides tagged with an 8-plex iTRAQ label in a MALDI-TOF/TOF mass spectrometer. Since ionization by the MALDI process almost exclusively leads to singly-charged peptides, irrespective of the number of basic residues in a peptide, most of the resulting peptides are of the immobile-proton (Arg at C-terminus) or partially mobile-proton type (Lys at C-terminus). In addition, the single charge of the peptide will usually be retained on the C-terminal fragment (*y*-ion) in case of tryptic peptides. A common observation in peptide fragmentation analysis is that peptide bond cleavage adjacent to certain amino acid residues is much more prevalent than to others, in both ESI (multiply charged) and MALDI (singly charged) derived peptides. For instance, proline uniquely has a secondary amine-group that strongly favors fragmentation of the peptide bond on its N-terminal side,[19,20] and acidic residues such as aspartic acid on their C-terminal side.[21,22]

The iTRAQ reagent is reactive towards amine groups and therefore leads to chemical modification of the N-terminal peptide amine and lysine residues.[23] The iTRAQ reagent is basic due to the presence of tertiary amine groups, and the reagent is therefore expected to influence the fragmentation pathway of a peptide, both by affecting proton mobility and the relative occurrence of *b*- and *y*-ions. The basic iTRAQ-modified N-terminal residue competes with Arg or Lys at the C-terminus for the proton, leading to spectra with a richer *b*-ion series compared with non-iTRAQ-labeled peptides.

This study analyzes the fragmentation pattern and relative fragment intensities of tryptic peptides by MALDI-TOF/TOF, in particular the effect of the iTRAQ label which is commonly used for quantification. Peptide identification based on intensity classification in decision trees is presented as a tool to improve peptide assignments.

2. Materials and Methods

2.1 Dataset preprocessing

Experimental datasets

Full proteome samples from two species, *Lactococcus lactis* and *Arabidopsis thaliana,* unlabeled or labeled with 8-plex iTRAQ (Applied Biosystems, Foster City, CA, USA), analyzed by 1D or 2D-LC-MALDI-TOF/TOF (for instrumental details see Gandhi et al., 2010) [24] were used in this study. MS/MS spectra were identified with Mascot 2.1 using species-specific databases.[25] The dataset labels and the number of

identified spectra and fragments (assigned *b*- and *y*-ions after spectra filtering) are listed in Table 1.

Spectra filtering

Singly-charged tryptic peptides with up to one missed cleavage, a Mascot significance score of greater than 99%, and a Mascot rank of 1 were selected across all datasets. Peptides identified with variable modifications were excluded. The spectra were preprocessed in the following manner: (1) fragment peaks with masses associated with the breakdown of the iTRAQ tag (reporter ions) were removed from each spectrum; (2) spectra with a goodness ratio of less than 0.25 were removed, where the goodness ratio is defined as the ratio of the sum of intensity of identified peaks (*b*- and *y*-ions) over the total intensity. This removes most mixed and low quality spectra; (3) spectra with ambiguous fragment identification arising from overlapping *b*- and *y*-ion masses were removed; (4) redundant fragments were removed where redundancy is defined as fragments with both the same peptide sequence and relative intensity (quartile rank, see below). The number of spectra and fragments post-filtering is shown in Table 1. Peptides with C-terminal arginine (immobile-proton) and lysine (partially mobile-proton) were separated in two different sets for each dataset.

Table 1. Experimental datasets used in this study

Dataset	Species	iTRAQ label	Spectra	Fragments
1	*A. thaliana*	None	2,451	66,454
2	*A. thaliana*	8-plex	5,222	107,919
3	*L. lactis*	8-plex	1,1076	195,867

Quartile ranks

The complexity of a spectrum was reduced by classifying its peaks in terms of quartile rank. The quartile rank is calculated by normalizing all identified fragment peaks to the most intense identified fragment in a given spectrum. The latter's intensity is divided in four equal parts to get the quartile intensity (QI) and quartile ranks are assigned to all fragments in the given spectrum, with rank 1 (Q1) assigned to the weakest peaks and rank 4 (Q4) assigned to the strongest. Sequence ions of the *b*- and y-series that are not found in the given spectrum are assumed to have intensity below the

signal-to-noise ratio cut-off. They are treated separately with a rank of 0 for the iTRAQ-labeled vs. unlabeled (non-iTRAQ) comparison and as first quartile (Q1) for the machine-learning analysis.

2.2 iTRAQ versus non-iTRAQ comparison

The quartile intensity distributions of peptide fragments from the *A. thaliana* sample, with (dataset 1) and without (dataset 2) iTRAQ labeling, were used to evaluate the effect of iTRAQ labeling on peptide fragmentation. This was done by creating pairwise fragmentation maps of abundances of peptide bond cleavage at each residue combination, one for each ion type (*b*- and *y*-ions), by plotting the respective average quartile rank. For the *b*-ions, fragments without any histidine, lysine, or arginine were used to avoid any secondary basic residue effect. All of the fragmentation maps were created using the same scale, with white indicating preference for missing peptide (rank 0) and black indicating preference for quartile rank of 4. Some of the rarely occurring amino acid residues, namely, cysteine, methionine, and tryptophan, were not displayed for clarity.

2.3 Machine learning-based classification

Decision-tree based approach

iTRAQ labeled peptides were classified by their quartile ranks with a decision-tree based learning approach, previously used with ESI data.[9] The training data set was constructed using *L. lactis*-based dataset 3. A total of 24 different sequence related features were calculated for each fragment in the training set, out of which 16 were found to be significant in the decision tree (Table 2). The tree was constructed by using the C4.5 algorithm with the pruning confidence level set at 95% and a minimum number of 200 cases required for a branch split.[26] C4.5-ofai (version 1.1) was used to print the generated pruned tree with a verbose setting. For this analysis, missing peptides were ranked as quartile 1.

Feature selection

Sequence related features known to affect fragmentation from past studies were used in our analysis.[9-12] Preference was given to features with values that are easy to interpret, such as true/false or labeled category. For instance, instead of using gas-based basicity values of residues, they were categorized as acidic, basic, or neutral. In general, the attributes are related to residue distance/length (e.g. distance to C-terminal), type of

residues immediate to the fragmentation site, and presence of internal residues (e.g. presence of Asp). Table 2 shows the features which were used by the decision tree for prediction of a fragment's quartile rank.

Match and Mismatch sets

In order to evaluate the decision tree, two test datasets were constructed, namely Match and Mismatch. The Match set was constructed with spectra from the *A. thaliana* dataset (dataset 2). For the Mismatch set spectra from *L. lactis* and *A. thaliana* (dataset 2 and 3) with a Mascot rank of 2 were selected, where the associated rank 1 peptide is identified with at least 99% confidence. The decision tree was used as input for a Java program to predict the quartile rank of the fragments from the Match and Mismatch datasets.

Scoring system

The quartile rank distribution at the end of each branch in the decision tree was used as the probability distribution for that stem. For scoring the QI of the fragments in a spectrum from the test datasets, the probability distribution from the appropriate stem is employed. Firstly, the quartile ranks are calculated for each spectrum in the dataset as described before. Secondly, a Quartile Intensity Score (QIS) is calculated for each spectrum by taking the sum of the probability of observing each of the quartile ranks. Thirdly, a theoretical maximum QIS (maxQIS) is also calculated by taking the sum of probability of the most likely quartile rank according to the decision tree for each of the fragments. If the observed spectrum perfectly matches its prediction, then the QIS would be equal to its maximum value, maxQIS. Finally, the QIS is normalized to maxQIS in the following manner: Normalized QIS = (QIS + ((1- maxQIS) x QIS)) x 100.

Table 2. Peptide sequence related features used by the decision tree

Symbol	Attribute	Values
DISTC	distance in amino acids to C-terminus	Continuous
DISTB	distance in amino acids to a basic residue C-terminal side of fragmentation site	Continuous
LENP	length of the peptide	Continuous

LENF	length of the fragment	Continuous
NUM_P	number of proline in the peptide	Continuous
NUM_D	number of aspartic acids in the peptide	Continuous
FRN_0	amino acid residue adjacent to the fragmentation site on N-terminal side	All amino acid residues
FRC_0	amino acid residue adjacent to the fragmentation site on C-terminal side	All amino acid residues
HIST_CF	presence of histidine in the charged fragment	True, false
HIST_UNCF	presence of histidine in the uncharged fragment	True, false
BRCH_UNCF	presence of branched chain amino acids in the uncharged fragment	True, false
PHN_0	FRN_0: acidic, basic, or neutral	A, B, N
PHC_0	FRC_0: acidic, basic, or neutral	A, B, N
ION_TYPE	ion type	b, y
GRN_0	FRN_0, grouped according to residue properties	Amide (Am), Aromatic (Ar), Small Hydrophillic (Sh), Large Hydrophobic (Lh), Small (sm), Acidic (A), Basic (B), Pro (P)
GRC_0	FRC_0, grouped according to residue properties	same as above

3. Results and discussion

3.1 Comparison of iTRAQ and non-iTRAQ datasets

Lysine terminated peptides

Pairwise fragmentation maps of average quartile rank for each peptide bond cleavage residue combination were created from a non-iTRAQ and 8-plex iTRAQ labeled sample, datasets 1 and 2 respectively. Figure 1 shows the fragmentation maps for Lys terminated peptides from the two sets, one each for *b*- and *y*-ions. Generally, the *y*-ions are observed as more intense peaks (QI4) than *b*-ions in both the labeled and the non-labeled set of peptides. However, in the *b*-ion maps fragments from the iTRAQ set are more intense than the ones from the non-labeled set. A closer look at the $Lys_{\{b,\ 8plex\}}$ map reveals an enhancement of bond cleavage with a Lys or Pro residue immediately C-terminal (FRC_0), or an Asp residue immediately N-terminal (FRN_0) to the fragmentation site.

Similarly, within the *y*-ion maps, Lys at FRC_0 and Asp at FRN_0 in the labeled peptides are enhanced over their non-labeled counterpart. This observation is in line with the known preferential fragmentation at acidic and proline residues. The effect of Lys is probably also related to fragment length, since by definition a Lys at FRC_0 is the C-terminal residue. The difference between iTRAQ and non-iTRAQ, however, is unexpected: for both *b*- and *y*-ions there is a positive effect if an 8-plex iTRAQ label is present. Since, lysine itself also has an iTRAQ label on its side chain amine, this may influence peptide fragmentation.

Arginine terminated peptides

Figure 2 shows the fragmentation maps of arginine terminated peptides. As expected, the *y*-ion maps show a selective fragmentation driven by an aspartic acid-based charge-remote pathway. The selectivity for this pathway is so strong that the presence of an iTRAQ label has no clearly visible effect on the *y*-ion fragmentation maps. However, within the *b*-ions, there is an enhancement for Pro at the FRC_0 residue. A more subtle enhancement is also seen for Asp at FRN_0, particularly at the residue combination of Asp-Arg (DR).

While the *b*-ions from the $Arg_{\{b,\ 8\text{-}plex\}}$ set are more abundant than their non-labeled counterparts, it is the $Lys_{\{b,\ 8\text{-}plex\}}$ map that displays a far richer *b*-ion set of peaks. This is likely to be a result of the modification of the N-terminal residue leading to two different events: 1) Asp driven charge-remote fragmentation with the proton sequestered at the N-terminus, and, 2) improved overall proton affinity on the N-terminal fragment (*b*-ion) in a charge-directed fragmentation. Both of these events play a

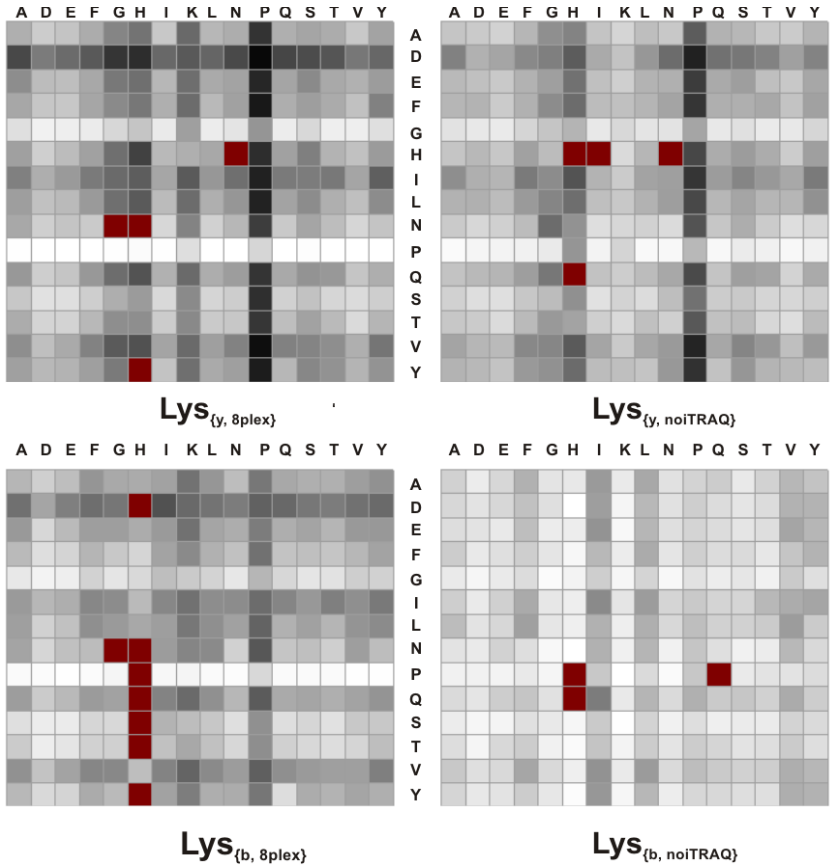


Figure 1. Pairwise fragmentation maps of lysine (Lys) terminated peptides.

The average quartile rank for bond cleavage at different residue combinations is shown by the hue, ranging from white (missed peptide) to black (quartile rank 4). The red hue indicates that less than 20 fragments were present in the dataset for that particular combination. The residues on the x-axis represent FRC_0 and those on the y-axis represent FRN_0. The maps on the top row stem from *y*-ions, whereas the ones on the bottom from *b*-ions. The maps in the left column are from peptides with an 8-plex *i*TRAQ label, whereas the maps in the right column consist of peptides without *i*TRAQ.

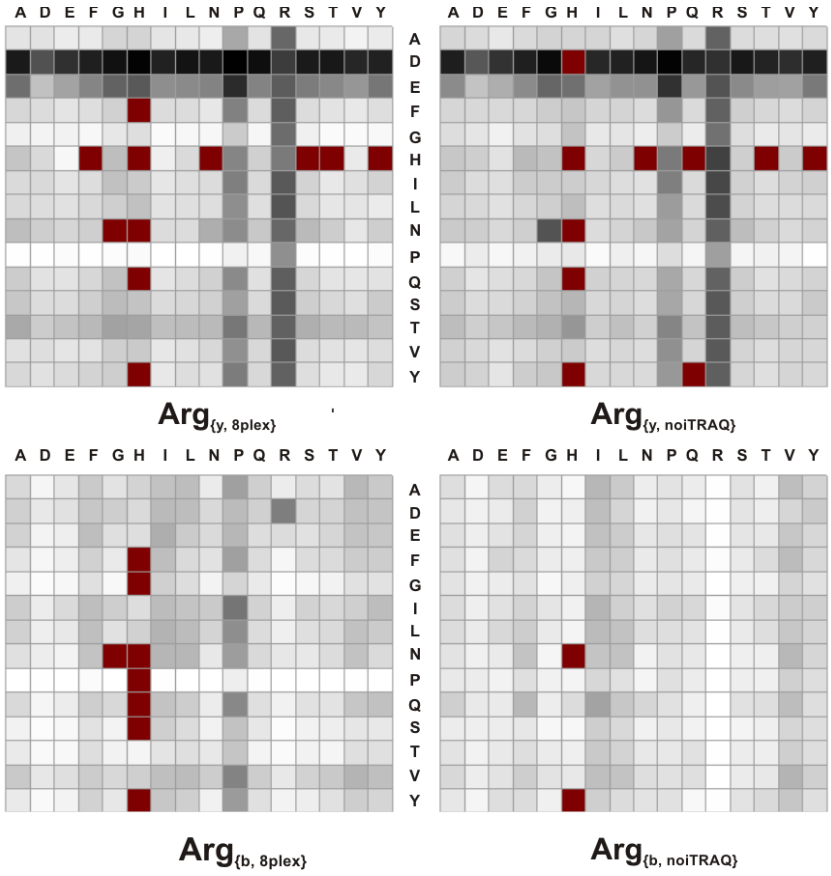


Figure 2. Pairwise fragmentation maps of arginine (Arg) terminating peptides. The average quartile rank for bond cleavage at different residue combinations is shown as detailed in figure 1.

more prominent role when there is a Lys at the C-terminus, instead of Arg which is a more basic residue. While elucidating the mechanisms behind the observed changes is outside the scope of this work, it is clear that iTRAQ labeling significantly influences peptide fragmentation.

3.2 Fragmentation model of iTRAQ modified peptides

In order to gain a deeper insight in the fragmentation pathways and factors influencing fragment intensity than is possible with heat maps, a fragmentation model of iTRAQ modified tryptic peptides was created using a decision-tree based machine-

learning approach, as described in the Materials and Methods section. The general tree consists of two distinct branches as expected, depending on whether the C-terminus is Lys, or Arg. Arginine and lysine-terminated peptides represent the vast majority of all identified tryptic peptides and their respective branches are discussed separately as the Arginine (Figure 3) and Lysine model (Figure 5). As was already expected from the fragmentation maps, the Lysine model is more complex in terms of number of nodes than the Arginine one. This is largely because of the strong selectivity for a few specific pathways present in the Arginine model. Both of the models were evaluated using test datasets to show the potential of using fragment ion abundance for predicting experimental spectra, and hence, aid with peptide and protein identification.

Description of Arginine model

In accordance with the fragmentation map (Figure 2), the prominent attribute in the Arginine model is the presence of an Asp residue (NUM_D). The tight sequestering of the proton at the Arg-terminus, favoring a charge-remote fragmentation next to Asp, leads to intense peaks in a spectrum. This "aspartic acid effect" forces an extreme distribution of the peak intensities: fragmentation events with a positive Asp effect are primarily Q4 (red), whereas those lacking in it (in the same peptide) are primarily Q1 (light blue). The strength of the aspartic effect is such that even glutamic acid (Glu) has little secondary impact on the relative intensity of the fragment. This effect is also seen in the *b*-ions when there is a His present in the fragment and the distance from the C-terminal arginine is 3 residues or less.

The patterns are more varied on the left side of the decision tree, i.e. the part of the tree without Asp playing a role. A key difference is the impact of the Glu residue in the absence of Asp, now behaving much like its more acidic counterpart. As seen in the Arg fragmentation map, the presence of Pro at the C-terminal side of the fragmentation site also enhances the fragmentation. However, this is qualified by the condition that the length of the fragment is less than or equal to 7 amino acids. This is likely due to it being a charge-directed fragmentation which requires the proton to move from the site of protonation to the Pro residue. The enhanced fragmentation seen for Arg at the FRC_0 position in the fragmentation map can also be traced back to the tree. However, now it comes with the additional information that the single amino acid fragment (Arg) stems from a peptide with no Pro (NUM_P = 0), Asp (NUM_D = 0) or His (HIST_UNCF = false) residue.

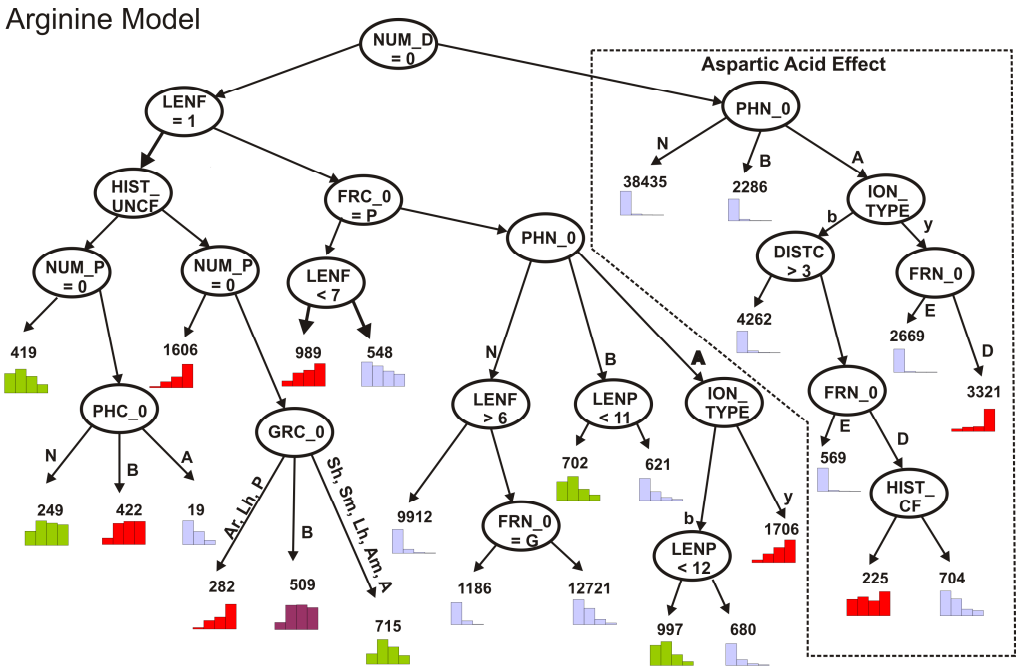


Figure 3. A graphical representation of the part of the decision tree containing arginine-terminated peptides (Arginine Model).

Labels in circles and along the arrows are attributes and attribute values, respectively, as listed in Table 1. The unlabelled arrows represent true (left arrow) or false (right arrow) for the corresponding condition. Histogram plots show the distribution of peptides over intensity quartiles Q4 (highest, left side) to Q1 (lowest, right side), with the number above it the total number of peptides in this branch. The colors indicate the maximum quartile rank in the given distribution (red for Q4, purple for Q3, green for Q2, and blue for Q1).

Prediction power of Arginine model

The Arginine Model was used to predict the relative intensities of all *b*- and *y*-ions in the Match and Mismatch test dataset (Figure 4A). Using a score cut-off of 80%, the model was able to correctly predict relative fragment intensities of 70% (1858) of the spectra in Match dataset. 264 of these 1858 spectra are predicted perfectly (all observed ions in the predicted quartile) by the model. Conversely, it also predicted 29% (750) of

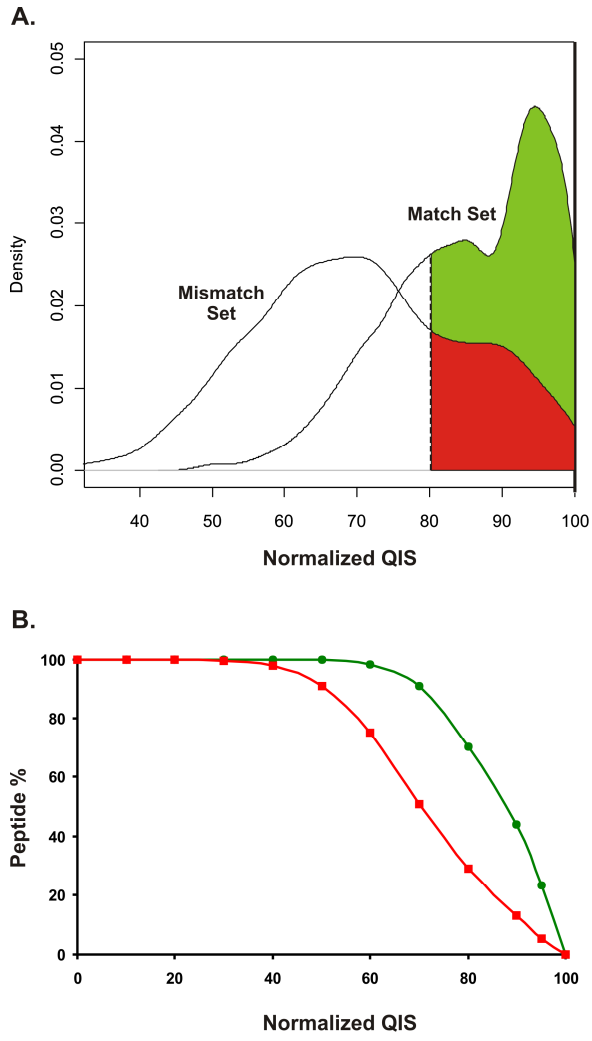


Figure 4. Performance of the Arginine Model when predicting the relative fragment intensities of peptides from Match and Mismatch datasets.

Panel A shows the density plot made from the histogram of peptide scores from Arginine Match and Mismatch datasets. The x-axis shows the normalized quartile intensity score for predicted peptides from the given dataset and the y-axis corresponds to the density of the number of peptides observed. The green area corresponds to the true positive matches and the red to the false positive matches with a score cut-off of 80%.

Panel B The x-axis shows the normalized quartile intensity score (QIS) for predicted peptides from the given dataset and the y-axis shows the corresponding percentage of the peptides with the same or higher normalized QIS. The red curve corresponds to the Mismatch dataset (True positives), whereas the green curve corresponds to the Match dataset (False positives).

the spectra in the Mismatch dataset (Figure 4B). While these results imply a high False Positive Rate for the model, this should be taken within the context of the likely presence of mixed peptide spectra in the Mismatch dataset. In addition, the peptide sequences in the Mismatch dataset often differ from the correct sequence by as few as a single amino acid residue. The model would not necessarily have the power to differentiate between two very close peptide sequences and hence, leads to an apparent false positive.

Description of Lysine model

In the absence of Arg residues with the ability to tightly sequester a proton, the charge-directed fragmentation events play a prominent role in the Lysine model. As seen in the $Lys_{\{y,\ 8\text{-plex}\}}$ fragmentation map, the primary attribute for Lys terminated peptides is the presence of a Pro at the C-terminal side of the fragmentation site (FRC_0). The presence of Proline at this site and observation of the *y*-ion leads almost exclusively to a Q4 fragmentation event which is dubbed the "Proline effect" in our model. As seen in the $Lys_{\{b,\ 8\text{-plex}\}}$ map, the proline effect is also found with *b*-ions in the decision tree. The decision tree model adds additional context to this observation, as follows: 1) lack of a His residue in the uncharged fragment, 2) distance of greater than 4 amino acids from the C-terminal basic Lys residue, 3) only a single proline residue in the peptide, and 4) a peptide length shorter than 22 amino acids. These conditions can largely be explained as factors where the proton affinity of the C-terminus fragment (*y*-ion) would be higher than of the N-terminal fragment (*b*-ion). An interesting suppressive effect stems from the presence of Glycine at the FRN_0 position. At this position, Glycine leads to a Q1 fragmentation event regardless of any other attributes, including a Pro at FRC_0 position. The Aspartic acid effect is still present in the Lysine model, albeit not as strongly as in the case of the Arginine model, since partially mobile protons are present in Lys-terminated peptides.

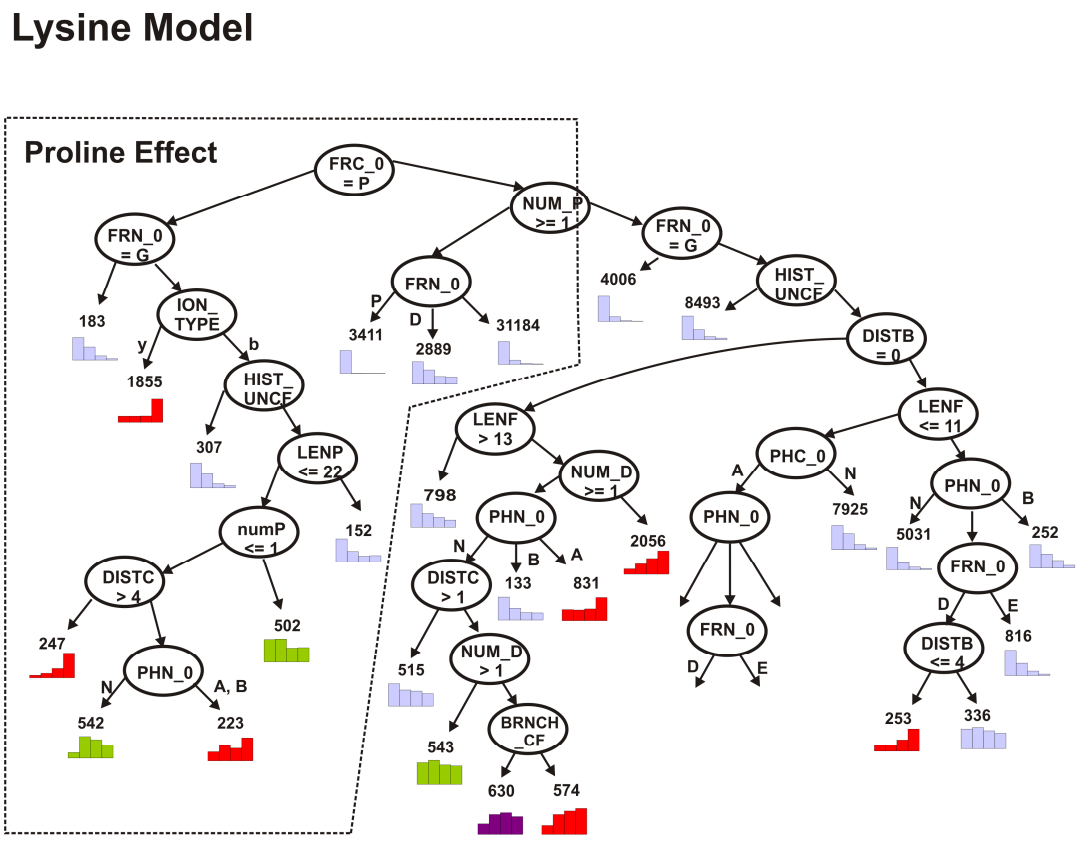


Figure 5. A graphical representation of the part of the decision tree containing lysine-terminated peptides (Lysine Model). See Figure 3 for the legend.

Prediction Power of Lysine model

The lysine model was used to predict the spectra in the lysine Match and Mismatch dataset. As shown in the density plot (Figure 6A), the scores from the Match set exhibit a mixed distribution. This is due to the difference in the prediction power of the model when predicting fragment intensity of peptides with or without a Pro residue. The model clearly performs much better when there is a Pro residue present in the peptide. When predicting the fragment intensity of peptides without a Pro residue, it only performs slightly better than the Mismatch scores. Using a score cut-off of 80, the model was able to predict 27% (695) of the 2585 spectra in Match dataset and 5% (62) of the 1303 spectra in the Mismatch dataset (Figure 6B). When only peptides with a Pro

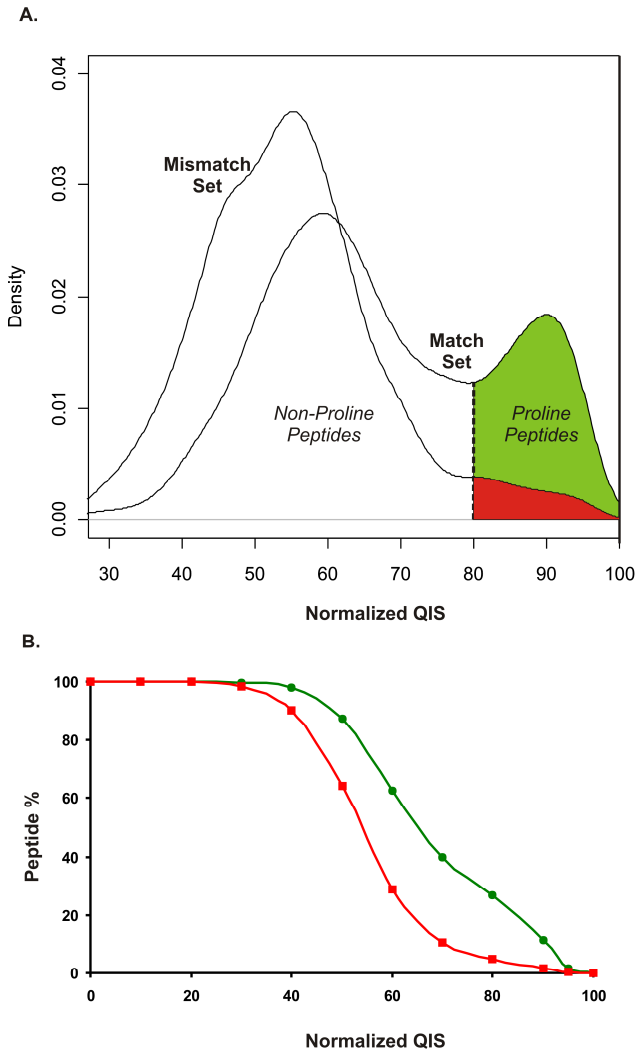


Figure 6. Performance of the Lysine Model when predicting relative fragment intensities of peptides from the Match and Mismatch datasets.

Panel A shows the density plot made from the histogram of peptide scores from Match and Mismatch datasets. The x-axis shows the normalized quartile intensity score for predicted peptides from the given dataset and the y-axis corresponds to the density of the number of peptides observed. The Match set shows two maxima dependent on the presence or absence of Pro in the peptide.

Panel B shows the percentage of peptides identified from the Match (blue curve) and Mismatch (red curve) dataset on the y-axis for the corresponding normalized QIS (x-axis).

residue are considered, it predicted 67% (689) of the 1027 spectra in the Match dataset and 24% (60) of the 254 spectra in the Mismatch dataset.

4. Conclusion

As shown here, iTRAQ modification has a significant influence on peptide fragmentation of singly charged peptides produced by MALDI. The presence of an additional basic group on the N-terminus leads to a more pronounced set of *b*-ion peaks. While all the factors involved in a fragmentation pathway are far from being known, probabilistic models can be made based on empirical observations. As more discoveries are made, these models can be refined further. We performed a rather simple intensity prediction by reducing the measured intensity, a continuous variable, into quartiles relative to the most abundant ion. Despite this simplification, we were able to show that the relative abundances of the ions can be predicted. While mass (m/z) information of an MS/MS spectrum will continue to play a pivotal role in peptide identification, there is enough evidence available to suggest that ion abundance in the form of peak intensity should also be used as another ingredient by search programs. However, ion abundance is heavily dependent on various factors such as instrumental set up, chemical modifications such as iTRAQ, and choice of digestion enzyme. Any protein search program that successfully utilizes intensity-based information will also need to account for such variations.

5. Acknowledgement

We thank F. Fusetti, E. Weiderhold, and P. Puri for providing with the MS datasets. This research work was supported by the Netherlands Bioinformatics Centre (NBIC) and the Netherlands Proteomics Centre (NPC).

6. References

1. Brunner, E. Ahrens, C. H. Mohanty, S. Baetschmann, H. Loevenich, S. Potthast, F. Deutsch, E. W. Panse, C. de Lichtenberg, U. Rinner, O. Lee, H. Pedrioli, P. G. A. Malmstrom, J. Koehler, K. Schrimpf, S. Krijgsveld, J. Kregenow, F. Heck, A. J. R. Hafen, E. Schlapbach, R. Aebersold, R. A high-quality catalog of the *Drosophila melanogaster* proteome. ***Nat. Biotechnol.*** 2007, 25, 576–583.
2. de Godoy, L. M. F. Olsen, J. V. de Souza, G. A. Li, G. Mortensen, P Mann, M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. ***Genome Biol.*** 2006, 7, R50.
3. Nesvizhskii, A. I. Roos, F. F. Grossmann, J. Vogelzang, M. Eddes, J. S. Gruissem, W. Baginsky, S. Aebersold, R., Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data. ***Mol. Cell. Proteomics*** 2006, 5, (4), 652–670.
4. Ning, K. Fermin, D. Nesvizhskii, A. I., Computational analysis of unassigned high–quality MS/MS spectra in proteomic data sets. ***Proteomics*** 2010, 10, (14), 2712–2718.
5. Kapp, E. A. Schütz, F. Connolly, L. M. Chakel, J. A. Meza, J. E. Miller, C. A. Fenyo, D. Eng, J. K. Adkins, J. N. Omenn, G. S. Simpson, R. J., An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. ***Proteomics*** 2005, 5, (13), 3475–3490.
6. Elias, J. E. Haas, W. Faherty, B. K. Gygi, S. P., Comparative evaluation of mass spectrometry platforms used in large–scale proteomics investigations. ***Nat. Meth.*** 2005, 2, (9), 667–675.
7. Bell, A. W. Deutsch, E. W. Au, C. E. Kearney, R. E. Beavis, R. Sechi, S. Nilsson, T. Bergeron, J. J. M., A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. ***Nat. Meth.*** 2009, 6, (6), 423–430.
8. Kapp, E. A., Schütz, F., Reid, G. E., Eddes, J. S. et al., ***Anal. Chem.*** 2003, 75, 6251–6264.
9. Elias, J. E. Gibbons, F. D. King, O. D. Roth, F. P. Gygi, S. P., Intensity-based protein identification by machine learning from a library of tandem mass spectra. ***Nat Biotech*** 2004, 22, (2), 214–219.
10. Zhang, Z., Prediction of low-energy collision-induced dissociation spectra of peptides. ***Anal. Chem.*** 2004, 76, (14), 3908–3922.
11. Huang, Y. Triscari, J. M. Tseng, G. C. Pasa–Tolic, L. Lipton, M. S. Smith, R. D. Wysocki, V. H., Statistical characterization of the charge state and residue dependence of low–energy CID peptide dissociation patterns. ***Anal. Chem.*** 2005, 77, (18), 5800–5813.
12. Khatun, J. Ramkissoon, K. Giddings, M. C., Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. ***Anal. Chem.*** 2007, 79, (8), 3032–3040.
13. Barton, S. J. Richardson, S. Perkins, D. N. Bellahn, I. Bryant, T. N. Whittaker, J. C., using statistical models to identify factors that have a role in defining the abundance of ions produced by tandem MS. ***Anal. Chem.*** 2007, 79, (15), 5601–5607.

14. Frank, A. M., Predicting intensity ranks of peptide fragment ions. ***J. Proteome Res.*** 2009, 8, (5), 2226–2240.
15. Dongre, A. R. Jones, J. L. Somogyi, A. Wysocki, V. H., Influence of peptide composition, gas–phase basicity, and chemical modification on fragmentation efficiency: evidence for the mobile proton model. ***J. Am. Chem. Soc.*** 1996, 118, (35), 8365–8374.
16. Summerfield, S. G. Whiting, A. Gaskell, S. J., Intra-ionic interactions in electrosprayed peptide ions. ***Int. J. Mass Spectrom. Ion Process.*** 1997, 162, (1–3), 149–161.
17. Gu, C. Somogyi, A. Wysocki, V. H. Medzihradszky, K. F., Fragmentation of protonated oligopeptides XLDVLQ (X = L, H, K or R) by surface–induced dissociation: additional evidence for the "mobile proton" model. ***Anal. Chim. Acta*** 1999, 397, (1–3), 247–256.
18. Wysocki, V. H. Tsaprailis, G. Smith, L. L. Breci, L. A., Mobile and localized protons: a framework for understanding peptide dissociation. ***J. Mass Spectrom.*** 2000, 35, (12), 1399–1406.
19. Hunt, D. F. Yates, J. R. Shabanowitz, J. Winston, S. Hauer, C. R., Protein sequencing by tandem mass spectrometry. ***Proc. Natl. Acad. Sci. U. S. A.*** 1986, 83, (17), 6233–6237.
20. Tabb, D. L. Smith, L. L. Breci, L. A. Wysocki, V. H. Lin, D. Yates, J. R., Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. ***Anal. Chem.*** 2003, 75, (5), 1155–1163.
21. Martin, R. L. Brancia, F. L., Analysis of high mass peptides using a novel matrix-assisted laser desorption/ionisation quadrupole ion trap time-of-flight mass spectrometer. ***Rapid Commun. Mass Spectrom.*** 2003, 17, (12), 1358–1365.
22. Tabb, D. L. Huang, Y. Wysocki, V. H. Yates, J. R., Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. ***Anal. Chem.*** 2004, 76, (5), 1243–1248.
23. Ross, P. L. Huang, Y. N. Marchese, J. N. Williamson, B. Parker, K. Hattan, S. Khainovski, N. Pillai, S. Dey, S. Daniels, S. Purkayastha, S. Juhasz, P. Martin, S. Bartlet-Jones, M. He, F. Jacobson, A. Pappin, D. J., Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine–reactive isobaric tagging reagents. ***Mol. Cell. Proteomics*** 2004, 3, (12), 1154–1169.
24. Gandhi, T. Fusetti, F. Wiederhold, E. Breitling, R. Poolman, B. Permentier, H. P., Apex peptide elution chain selection: a new strategy for selecting precursors in 2D–LC-MALDI-TOF/TOF experiments on complex biological samples. ***J. Proteome Res.*** 9, (11), 5922–5928.
25. Perkins, D. N. Probability-based protein identification by searching sequence databases using mass spectrometry data. ***Electrophoresis*** 1999, 20, 3551–3567.
26. Quinlan, J. R. C4.5: programs for machine learning. Morgan Kaufmann Publishers, 1993.

Chapter IV

Detecting significant protein enrichment in subtractive proteomics: quest to identify the yeast vacuolar membrane proteome

Parts of this chapter were published in *Mol. Cell. Proteomics 9,* 431-445 (**2010**)

Abstract

Subtractive proteomics is a powerful strategy for identifying organelle specific proteins. Since organelle preparations are never free of contaminants due to limitations in fractionation methods, employing a normal proteomics approach would lead to identifying a large number of false positives, in addition to the true organelle proteins. This limitation can be alleviated by subtracting proteins identified in a crude sample from those identified in an organelle enriched sample and thereby, allowing the identification of proteins truly localized to the organelle of interest. In this study, we perform subtractive proteomics to distinguish contaminants from true vacuolar proteins in the yeast *Saccharomyces cerevisiae* by comparing the relative abundances of proteins in pure and crude preparations. A robust statistical analysis was performed to identify vacuolar proteins by detecting significant changes in protein abundances.

1. Introduction

The vacuole is the largest organelle of yeast cells and the functional equivalent of the mammalian lysosome. The vacuole is surrounded by a single membrane, which contains the V-ATPase complex that acidifies the interior of the organelle. The pH difference between the vacuolar lumen and the cytosol is used as the driving force for transport across the vacuolar membrane.[1] These transport processes are important for many crucial functions of vacuoles: storage of organic molecules, detoxification, proton and ion homeostasis, and proteolysis of cytosolic and membrane proteins.[2-9]

Whereas vacuolar luminal proteins have been studied fairly extensively,[10,11] our knowledge about the integral membrane proteins in the yeast vacuolar membrane is limited. A handful of vacuolar transport proteins have been identified and characterized by classical genetic and biochemical approaches.[12-14] However, based on measurements of transport activities across the vacuolar membrane and determination of organic substance content in the vacuole lumen, the existence of many more proteins with translocation activities is expected.

Approximately 1000 yeast proteins (17–18% of the entire *S. cerevisiae* proteome) do not have an annotated localization in the Saccharomyces Genome Database (SGD; www.yeastgenome.org, release version 06.09.2007). It is likely that some of these are of vacuolar origin. In recent years a number of global protein localization studies[15-18] revealed the vacuolar localization of more than 40 putative proteins with unknown biological functions. However, these studies might have failed to detect low-abundant

membrane proteins, and possibly yielded incorrect information in some cases, as the tags (such as the C-terminal GFP tag) [15] may affect targeting, resulting in the mis-localization of proteins.

To characterize the vacuolar membrane proteome and to identify novel vacuolar membrane proteins, we produced highly purified vacuolar membranes and analyzed their protein content by mass spectrometry. To discriminate between genuine vacuolar residents and contaminating proteins the subtractive proteomics technique LOPIT (Localization of Organelle Proteins by Isotope Tagging) was used,[19-20] in conjunction with iTRAQ-based quantification. Proteomics experiments using iTRAQ labeling can determine the relative abundance of a large number of proteins under different conditions. The results are often presented as lists of differentially abundant proteins on the basis of iTRAQ quantification data. In order to analyze and validate these lists, we use a modified version of the statistical technique known as *"Iterative Group Analysis"* (iGA), which was originally developed for analyzing lists of differentially expressed genes found in microarray experiments.[21] This technique, which we called Double Boundary iGA (db-iGA), is especially suitable for detecting concerted changes in protein abundances due to the enrichment of vacuolar proteins. The db-iGA method resulted in the identification of a group of 148 proteins that was enriched along with known vacuolar proteins in our preparation. In this group, 22 proteins without annotated localization could be assigned as likely to be vacuolar, of which at least 9 have confirmed or predicted translocation activity.

2. Materials and Methods

2.1 Experimental Procedure

Yeast strain and cell growth

Haploid *Saccharomyces cerevisiae* W303 (MATα *ade2-1 leu2-3,112 his3-22,15 trp1-1 ura3-1 can1-100*) was used.[22] All experiments were carried out as biological quadruplicates. For cell growth, 10 ml of YPD medium (0.3% yeast extract, 0.5% Bacto-peptone, 1% glucose) was inoculated with a colony from a fresh agar plate, and incubated in a 100 mL Erlenmeyer flask for 15 hr at 30°C (shaking speed 160 rpm). 50 mL of fresh YPD medium was inoculated with 0.5 ml of the pre-culture. Cells were grown for 6–8 hr until the density was 1.5-2 x 10^7 cells/mL. At least 4 subsequent 200-500 fold dilutions in fresh medium followed by growth to a density of 1.5–2 x 10^7

cells/mL were carried out. Finally, for large-scale preparation, 12 L of medium in a pH-, pO_2- and temperature-controlled bioreactor was inoculated with 20 mL of the last pre-culture. Cells were grown aerobically (30% oxygen saturation, stirring speed 150 rpm) at a controlled pH of 6.3 in YPD medium. The doubling time was 1.5 hr. Exponentially growing cells (1.5 x 10^7 cells/mL) were harvested by centrifugation at 4000 x g_{avg} for 5 min. Unless indicated otherwise, all steps were performed at Room Temperature (RT). The cells were washed with 1 L double distilled water and centrifuged again. The wet weight of cells from a 12 L culture was 20–30 g.

Isolation of intact vacuoles and preparation of vacuolar membranes

Spheroplast formation was carried out according to Kipper *et al.*[23] with a few alterations, as specified in Wiederhold et al.[24] Light microscopy was used to evaluate the extent of spheroplasts formation. After digestion of the cell wall, the spheroplasts were cooled on ice and washed by centrifugation through a Ficoll-sorbitol layer. For isolation of intact vacuoles, the spheroplasts were lysed as described in the protocol of Ohsumi *et al.*,[25] with modifications detailed in Wiederhold et al.[24] The vacuoles – usually approximately 5 mL per 20 g of cells – were lysed osmotically and the vacuolar membranes were purified by differential centrifugation as described in Wiederhold et al.[24] The protein concentration was determined by the BCA method (Pierce) after solubilization in 2% SDS.

Sample preparation for SCX/RP-LC and iTRAQ labeling

For trypsinization, 100 µg of protein was resuspended in 30 µL of 500 mM TEAB, 33% methanol plus 0.05% SDS. Reduction of disulfide bonds with Tris(2-carboxyethyl) phosphine hydrochloride (TCEP), cysteine-modification with methyl-methanethiosulfonate (MMTS), digestion with trypsin (Cat.: V511A, Promega) and iTRAQ-labeling were performed according to the manufacturer's protocol (Applied Biosystems). For each biological replicate the peptides derived from proteins in the crude and pure vacuolar membrane preparations were labeled with two different iTRAQ reagents. For the first replicate iTRAQ reagents 114 (for the pure sample) and 116 (for the crude sample) were used; and for the second replicate reagents 115 (pure) and 117 (crude). Then, equal amounts of the four sets of labeled peptides from the two biological replicates were combined. For the third and fourth biological replicates a label swap was done: the iTRAQ reagents 114 and 115 were used for the crude samples and reagents 116 and 117 for the pure samples. The two peptide mixtures (combined

biological replicates 1 and 2; and combined biological replicates 3 and 4) were subjected to chromatography and mass spectrometry analysis

Pre-fractionation of peptides on SCX

For off-line peptide pre-fractionation, a silica-based Polysulfoethyl Aspartamide strong cation exchange (SCX) column was used (Cat.: 202SE0502 PolyLC Inc., Columbia USA). The column was run at a flow rate of 200 μL/min on an Ettan-MDLC system (Amersham Biosciences AB, Uppsala, Sweden). Gradient solutions A: 10 mM KH_2PO_4-H_3PO_4, pH 2.7, 25% acetonitrile (ACN); B: 10 mM KH_2PO_4-H_3PO_4, pH 2.7, 25% ACN, 1 M KCl. Gradient conditions: column equilibration with 5 column volumes (CV) (1 CV = 0.7 ml) of 100% A. Peptides were loaded in 100% A and the column was washed with 5 CV from 0 to 3% B, elution: 3 to 12% B in 12 CV, followed by 12 to 30% B in 3 CV. Fractions were collected every 30 sec in 96-well plates. Eluted peptides were concentrated to approximately 40 μL in a vacuum centrifuge and diluted 1:2 with 0.2% TFA. Depending on the complexity, either separate fractions or pools of two fractions were analyzed by RP-LC MALDI-TOF/TOF.

RP-LC and MALDI-TOF/TOF analysis

Peptides were trapped on a pre-column (Cat.: 5065-9914, Zorbax 300SB-C_{18}, Agilent Technologies, Santa Clara CA, US) and then separated on a 75 μm × 150 mm analytical column (Cat.: 5065-9911, Zorbax 300SB-C_{18}, Agilent Technologies) using the Ettan-MDLC nanoLC system in the high-throughput configuration (Amersham Biosciences AB, Uppsala, Sweden). Gradient solutions contained A: 0.065% TFA; B: 0.065% TFA, 84% ACN. Gradient conditions: equilibration of column, binding and washing of peptides was performed with 3% B, elution with 3 to 30% B in 60 min. The eluting peptides were mixed 1:4 with 2.2 mg/ml α-cyano-4-hydroxycinnamic acid matrix (LaserBio Labs, Sophia-Antipolis, France), 3 fmol/μL of Angiotensin II (Sigma Aldrich), and 6 fmol/μl of Adrenocorticotropic hormone (ACTH) fragment 18-39 (Sigma Aldrich) and spotted directly onto a MALDI target (200 spots), using a Probot system (LC Packings, Amsterdam, The Netherlands). Peptides were analyzed with a 4700 Proteomics analyzer (Applied Biosystems, Foster City, CA, USA) MALDI-TOF/TOF mass spectrometer.

The MALDI-TOF/TOF was operated in reflectron-positive ionization mode in the m/z range 900-5000. The 30 most intense peaks above the signal-to-noise (S/N)

threshold of 150 from each MS spectrum were selected for MS/MS fragmentation and in addition the 10 most intense peaks in the S/N range of 150-30 were selected from same MS spectrum. The MS/MS spectra were acquired using 1 kV acceleration voltage and air as collision gas at 5 x 10^{-7} Torr. The precursor mass transmission window was set to ± 5 Da. The peak-lists of the acquired MS/MS spectra were generated, using default settings and a S/N threshold of 10. The MS spectra were calibrated internally, using Angiotensin II (m/z = 1046.542) and ACTH 18-39 (m/z = 2465.199). MS/MS calibration of the instrument was performed daily, using ACTH 18-39 fragment ions.

2.2 Criteria for protein identification and quantification

MS/MS peak-lists were extracted by GPS Explorer Software, version 3.5 (Applied Biosystems), using default parameters and were automatically submitted to a database search. All MS/MS spectra were analyzed using Mascot (Matrix Science, London, UK; version 1.9.05) and X! Tandem (www.thegpm.org; version 2006.04.01.2). Mascot and X! Tandem were set up to search a combined *S. cerevisiae* Genome Database (SGD), assuming the digestion by trypsin and allowing one or two missed cleavages for Mascot or X! Tandem, respectively. The database was created by combining forward and reversed entries of the SGD (release version 06.09.2007) and included sequences of porcine trypsin (NCBI accession: P00761) and human keratins (P35908, P35527, P13645, NP_006112) containing in total 13,443 protein entries. Mascot and X! Tandem were searched with a fragment ion mass tolerance of 0.20 Da and a parent ion tolerance of 200 ppm. MMTS modification of cysteine and Applied Biosystems iTRAQ multiplexed quantitation chemistry of lysine and the N-terminus were specified in Mascot and X! Tandem as fixed modifications. Deamidation of asparagine and glutamine, oxidation of methionine and Applied Biosystems iTRAQ multiplexed quantitation chemistry of tyrosine were specified in Mascot and X! Tandem as variable modifications.

Scaffold (version Scaffold-01_06_17, Proteome Software Inc., Portland, OR) was used to validate MS/MS-based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0% probability as specified by the Peptide Prophet algorithm.[26] Protein identifications were accepted if they could be established at greater than 99.0% probability and contained at least 2 uniquely identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm.[27] Proteins that contained similar peptides and could not be

differentiated based on MS/MS analysis alone were grouped to satisfy the principle of parsimony. Those peptides were removed from the dataset when quantification was performed. The false positive rate was calculated by dividing 2 times the number of proteins identified in the reversed database by 13,443, the sum of all proteins identified in forward and reversed versions of SGD. In two biological replicates, the false positive rate was 0.0035% and in two other biological replicates no hits from the reversed database were detected, using the criteria described above.

The relative quantification was based on peptides that were chemically labeled with isobaric reagents, using the iTRAQ technique. The quantification information was obtained automatically by GPS Explorer software from the peak areas of the reporter ions (m/z 114.1, 115.1, 116.1 and 117.1, with a mass tolerance of 0.1 Da) from the MS/MS spectra. The peak areas were corrected for isotopic impurities by the GPS Explorer using the information provided by the manufacturer in the Certificate of Analysis for each iTRAQ-multiplex batch. Each protein quantification was based on two or four biological replicates and if it was based on a single peptide only in one replicate, at least two identified peptides were necessary in one of the other biological replicates for a protein to be included in the analysis. Peptides that matched to multiple proteins were excluded from quantification, and indistinguishable isoforms of the same protein (e.g. ribosomal subunits A and B) were reported without the isoform specification. To select quantification data, those ratios were removed where the peak area of one reporter ion was below the signal-to-noise threshold of 10.

2.3 Statistical analysis

Enrichment Ranking

Enrichment ranking was based on the Rank Products statistics,[28] which was modified for use on peptide data as follows. If the same peptide was measured multiple times in the same biological sample, the iTRAQ ratios were averaged by calculating the mean. The peptide iTRAQ ratios were then listed in descending order and the ranks were assigned so that the peptide with the highest ratio had rank 1, the peptide with the second highest ratio had rank 2, and so on. In this way, *peptide ranks* were obtained. The ranks of peptides derived from the same protein were averaged by calculating the median to minimize the effect of outliers and the resulting medians were then ranked again, resulting in *protein ranks* for each biological replicate. To combine the protein ranks of all replicates, the median of protein ranks across replicates was calculated and

subsequently ranked again. This resulted in a single list of proteins, with the most consistently enriched protein at the top (rank 1).

Double-boundary Iterative Group Analysis

To determine which proteins were co-enriched with known vacuolar proteins, a modified version of the iterative Group Analysis algorithm (iGA) was applied.[21] The iGA method is rooted in the hypergeometric distribution, which describes the number of successes in a sequence of *n* draws from a finite population without replacement. In terms of a ranked list of annotated proteins (finite population), "number of successes" can be described as observing proteins belonging to a specific group, e.g. vacuolar proteins, within a specific window (*n* draws) of the list. The method simply asks the question of how likely it is to observe *"this many"* proteins belonging to a specific group "*that*" high up in the list *by chance*. In order to perform this type of analysis, proteins were assigned to one or several subcellular localizations, based on SGD annotations (release version 06.09.2007). "Cytosolic" and "cytoplasmic" localizations were combined, as were "nucleus", "nucleolus", "nuclear membrane", and "nuclear pore". "Unknown" and unspecified "membrane" proteins were also combined in a single class. In cases of multiple annotations (e.g. plasma membrane and vacuole), multiple classes were assigned and used for cluster analysis by iGA. We extend the original iGA method, which only analyses enrichment at the top, by iterating instead over both the start and end position within the ranked list. For each localization class the list of enrichment-ranked proteins was analyzed using all possible windows to define groups. The window that showed the most surprising clustering of proteins from the same localization class – which is the highest probability of change [–log(PC)] value as defined in Breitling *et al.*[28] – was recorded. This procedure, which was called double-boundary iGA, is more flexible than the original iGA approach, which only tests windows at the extremes (top or bottom) of the list and would for example miss clustering in the middle. Multiple testing corrected p-values were determined using 1,000 random permutations of the protein list for each class. Corrected p-values <0.005 were regarded as significant. The method was implemented using Wolfram Mathematica (version 6.0). Besides adding the functionality of db-iGA, our implementation is also more reliable than the one originally written in Perl, due to better handling of large integer calculations. The program creates two different types of output: (1) a text output with the list of all PC values, and (2) plots of the values in 2D (for iGA) and 3D (for db-iGA). The db-iGA method is

available for download from the NBIC development project environment website (https://trac.nbic.nl/db-iga/downloads).

3. Results

3.1 Subtractive proteomics of purified vacuoles

For the proteomic analysis of yeast vacuolar membrane proteins, it was necessary to isolate vacuoles of a high purity. The protocol of Ohsumi *et al.*[25] was optimized and used in combination with the spheroplasting procedure from Kipper *et al.*[23] As detailed in the Experimental section, we used two density centrifugation steps. In both steps the vacuoles floated to the top of the tube. The first centrifugation step yielded crude vacuoles, and after the second step we obtained highly pure vacuoles.

To be able to distinguish vacuolar proteins from contaminants, we used the subtractive proteomics technique LOPIT (Localization of Organelle Proteins by Isotope Tagging). The idea of LOPIT is that the set of proteins that is physically associated with an organelle will co-enrich during the organelle isolation, whereas contaminants – although still present – will be depleted. We compared the relative abundances of proteins in the purified vacuoles (obtained after the second density centrifugation step) and the crude vacuoles (obtained after the first density centrifugation step). The strategy is outlined in Figure 1.

To facilitate the identification of low-abundant vacuolar membrane proteins in the mass spectrometry analysis, vacuolar membranes were prepared in two steps to remove luminal and peripheral proteins. In the first step, vacuoles were lysed in the presence of EDTA and vacuolar membranes were spun down by ultracentrifugation, and in the second step the membranes were stripped with sodium carbonate (pH 11.8).
In each of the four biological replicates the relative proteins abundances of the pure and crude vacuolar membranes were compared. The isobaric 4-plex iTRAQ reagents were used to differentially label tryptic peptides derived from proteins in the crude and pure vacuolar membrane preparations. Because four different iTRAQ labels were available (114/115/116/117), we could combine the pairs of labeled peptides from two different biological replicates for subsequent chromatographic separation and mass spectrometry analysis: peptides from crude vacuolar membranes were labeled with iTRAQ reagents 116 (replicate 1) or 117 (replicate 2), and those derived from the pure membranes were labeled with iTRAQ reagents 114 (replicate 1) or 115 (replicate 2). In a second mixture, the other two biological replicates were combined and the labeling combinations were

reversed, i.e. peptides from crude vacuolar membranes were labeled with iTRAQ reagents 114 (replicate 3) or 115 (replicate 4) and those from pure vacuolar membranes with 116 or 117. The labeled peptides mixtures were fractionated using cation exchange and reversed-phase chromatography. Peptides were identified by tandem mass spectrometry (MS/MS). The reporter peaks of the iTRAQ reagents in the MS/MS spectra were used for quantification. The reporter peak area of a peptide derived from proteins present in the *pure* vacuolar membranes were divided by the reporter peak area of the peptide derived from proteins in the *crude* membranes, resulting in iTRAQ ratios. So, in theory, peptides with ratios larger than 1 were enriched in the pure vacuolar membrane fraction and, conversely, ratios smaller than 1 indicated depletion (peptides derived from contaminant proteins).

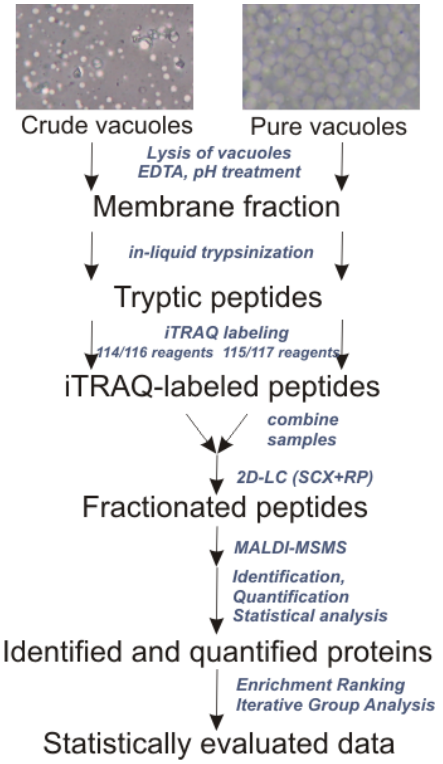


Figure 1. Workflow of LOPIT

To compare relative abundances of proteins, the membranes of the crude and pure vacuoles were stripped using EDTA and sodium carbonate at pH 11 and subsequently digested with trypsin in the presence of 30% methanol (v/v). The tryptic peptides were

labeled pair-wise with different 4-plex iTRAQ reagents and the labeled peptides derived from the crude (114/116 reagents) and pure (115/117 reagents) vacuolar membranes were combined. To reduce the complexity, the combined peptides were pre-fractionated on a strong cation exchanger (SCX) and each fraction was subjected to RP-LC-MALDI analysis. The acquired spectra were analyzed by Mascot and X!Tandem. The identified and quantified peptides were ranked according to their iTRAQ ratios and significant groups of enriched and depleted proteins were determined using double-boundary iGA.

572 proteins were identified and quantified, of which 484 (84%) were found in both mass spectrometry analyses. Because only those peptides were quantified for which all iTRAQ reporter fragments (114/115/116/117) were detected, the 484 proteins had been present in all four biological replicates and the remaining 88 proteins had been present in at least two biological replicates, indicating a high reproducibility of the procedure. As expected, the identified proteins were derived from various cellular locations, because a mixture of highly pure and crude vacuolar membranes (containing contaminants) was analyzed (Figure 2). Almost 60% of the 572 identified proteins

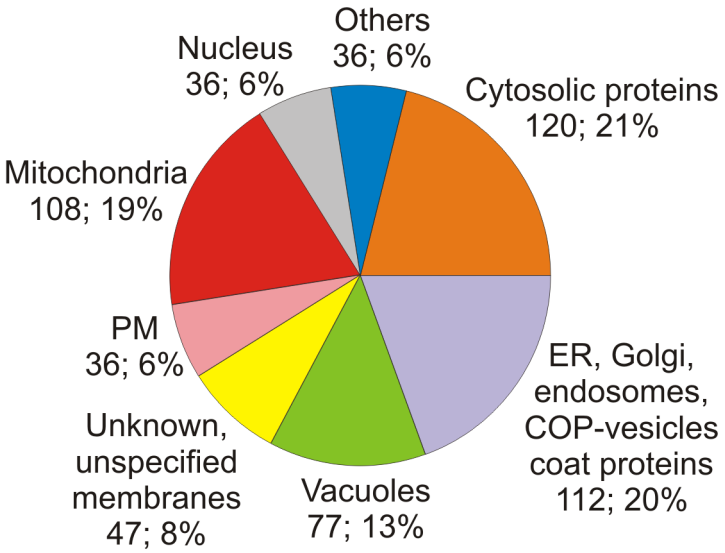


Figure 2. Annotated protein localizations of all identified and quantified proteins
The entire dataset contained 572 proteins and included both true vacuolar proteins and contaminants. The 77 identified vacuolar proteins cover 42% of all proteins annotated as "vacuolar" in SGD.

belonged to three major groups: cytosolic proteins, mitochondrial proteins, and components of the protein trafficking machinery including ER, Golgi, COP-vesicle coat proteins and the endosome. More than 63% (363 proteins) of the identified proteins did not contain predicted transmembrane domains. In the entire data set, 77 proteins were annotated "vacuolar" in SGD, representing approximately 42% of the total number of proteins annotated as vacuolar. The remaining 58% of annotated vacuolar proteins were not found and reasons for their absence will be discussed below. Intriguingly, 47 proteins of the 572 proteins identified in our analysis did not have a known localization. Such proteins with unknown localizations are candidate novel vacuolar (membrane) proteins.

3.2 Enrichment Ranking and iterative Group Analysis

In theory, each protein with an absolute iTRAQ ratio larger than 1 could be considered enriched in the pure vacuolar preparation. In practice, however, experimental errors prior to the mixing of the peptides such as those associated with protein concentration determination or incomplete digestion or labeling, interfere with such direct assignment. To reliably determine which of the identified proteins were enriched in the purified vacuolar fraction, Enrichment Ranking in conjunction with iterative Group Analysis (iGA) was applied.[21,28]

A list of all identified proteins was made, ranked in the order of descending iTRAQ ratios. To this end, the data from multiple peptides per protein and multiple biological replicates were analyzed in a robust way, as described in detail in Materials and Methods. We then used a statistical test, based on iGA, to determine if proteins from particular subcellular localizations as annotated in the SGD were significantly clustered in the list. iGA was originally developed for the functional annotation of microarray results and is based on hypergeometric statistics. It is applicable to small and noisy data sets with a relatively low number of replicates. A major advantage for the analysis is the robustness of iGA against imperfect assignments of the functional classes, in our case incorrect or incomplete annotation of the localization of proteins in the SGD. To make iGA applicable to our purposes, we modified the original algorithm as described in Experimental Procedure, resulting in the so-called double-boundary iGA (db-iGA) approach. We applied the double-boundary iGA to the set of 572 proteins that were ranked as described above.

As expected, known vacuolar proteins were clustered among the proteins with the highest iTRAQ ratios, i.e. the proteins that were most enriched in the density centrifugation. The iGA defined a group of 148 proteins at the top of the ranked list in which most annotated vacuolar proteins were clustered (PC value of 3.4×10^{-37}; multiple-testing corrected p-value < 0.001). This group was named the enriched cluster, and contained 69 proteins annotated as vacuolar, 22 proteins with no known localization and 57 proteins annotated as localized to other organelles (Figure 4A). The latter were mainly from the ER-Golgi-endosome network, cytosol and plasma membrane. These proteins, which are co-enriched with the vacuolar proteins, may represent proteins targeted to the vacuoles for degradation, proteins with multiple localizations, or proteins with incomplete or incorrect database annotation, and they will be discussed below. The 22 proteins with no known localization represent potential novel vacuolar proteins (Table 1). As mentioned above, in the entire data set we could identify 77 proteins annotated as vacuolar. The iGA analysis excluded 8 of these proteins from the enriched vacuolar cluster. Possible reasons for the exclusion will be discussed below.

Table 1 Novel Vacuolar Proteins

Twenty-two proteins without known localizations were determined as enriched along with pure vacuolar membranes (Figure 4). Nine proteins are predicted transporters and thirteen display versatile functions.

Protein accession number	**Protein name**	**TM***	**(Predicted) function****	**Localization ****	**GFP localization**
TRANSPORT PROTEINS					
YAL022C	FUN26	11	Nucleoside transporter	membrane	no
YBR235W	YBR235W	10	Cation/chloride co-transporter	unknown	ambiguous
YCR011C	ADP1	7	ABC-transporter	membrane	ER
YDL054C	MCH1	11	MFS transporter	membrane	vacuolar membrane
YGL114W	YGL114W	12	Oligopeptide transporter	membrane	No
YJR124C	YJR124C	9	MFS transporter	unknown	No

YKL064W	MNR2	2	Magnesium and cobalt ion transporter	membrane	Ambiguous
YLR047C	FRE8	5	Ferric reductase-like transmembrane component	membrane	No
YOR291W	YOR291W	11	Cation-transporter ATPase	membrane	ER
OTHER FUNCTIONS					
YAR028W	YAR028W	2	Unknown	unknown	Ambiguous
YBL050W	SEC17	0	Soluble NSF attachment protein (SNAP) involved in ER to Golgi transport	unknown	No
YBR074W	YBR074W	8	Metallo-endopeptidase	unknown	No
YDR089W	YDR089W	3	Protein involved in membrane organization and biogenesis	unknown	Ambiguous
YGR141W	VPS62	0	Protein involved in protein targeting to vacuole	unknown	No
YLR173W	YLR173W	1	Unknown	unknown	No
YLR240W	VPS34	0	Protein kinase	unknown	punctate, endosome
YLR241W	YLR241W	11	RSN (yeast)-related membrane protein	unknown	No
YLR360W	VPS38	0	Protein involved in late endosome to vacuole transport	unknown	Endosome
YMR266W	RSN1	11	Protein involved in Golgi to plasma	unknown	cell periphery

			membrane trafficking		
YOR034C	AKR2	7	Zn finger DHHC-domain containing protein involved in endocytosis	unknown	No
YPL057C	SUR1	2	Mannosyltransferase involved in sphingolipid biosynthesis	intracellular	vacuolar lumen
YPL120W	VPS30	0	Protein involved in targeting to vacuole and autophagy	membrane	vacuolar lumen

* Displays the number of transmembrane domains corrected for N-terminal signal peptide. From the number of transmembrane helices predicted by TMHMM, the first TMD was subtracted if it overlapped with the signal peptide as predicted by SignalP.
** Information was obtained from SGD

Table 2. Double-boundary iGA results

The double-boundary iGA yielded significant clustering for different groups of proteins for which the probability of change (PC-) values are shown. The significance of clustering was tested using a multiple-testing procedure; corrected p-values below 0.001 indicate high significance, corrected p-values above 0.005 were regarded as insignificant.

Group	Range	PC-value	Number of group members (Total) Above-**within**-below Boundaries	**Corrected p-value**
ABOVE SIGNIFICANCE THRESHOLD				
Vacuole	1–148	$3.4x10^{-37}$	(77) 0-**69**-8	**<0.001**
Endosome	88-214	$1.9x10^{-9}$	(13) 0-**13**-0	**<0.001**
Mitochondrion	326-570	$1.9x10^{-12}$	(127) 37-**89**-1	**<0.001**
BELOW SIGNIFICANCE THRESHOLD				
Cytosol	98-505	$1.5x10^{-5}$	(127) 8-**109**-10	**0.013**

ER	211-518	$9.7x10^{-6}$	(74) 14-**57**-3	**0.008**
Golgi	55–256	$6x10^{-4}$	(33) 1-**21**-11	**0.251**
Nucleus	279-381	$1.5x10^{-5}$	(38) 7-**8**-13	**0.017**
Plasma membrane	7-310	$4x10^{-4}$	(39) 0-**31**-8	**0.249**

3.3 Double-boundary iGA determined clusters of proteins

The double-boundary iGA was used to test whether proteins with different annotated localizations were significantly clustered in the ranked list. The ranges, PC-values and corrected p-values for each cluster are summarized in Table 2. The group containing mitochondrial proteins was well defined at the bottom of the list and was regarded as strongly depleted from the pure vacuolar fraction. The group of endosomal proteins was found to be significantly clustered, but it was overlapping with the cluster of enriched vacuolar proteins (Figure 3). The overlap suggested that some of the proteins were co-enriched with the vacuolar proteins and might represent proteins targeted to vacuoles for degradation or, more likely, proteins with multiple localizations, as will be discussed below. Proteins with other annotated localization (ER, Golgi, plasma membrane, nucleus, cytosol) did not form significant clusters.

4. Discussion

4.1 Proteins with annotated vacuolar localization

Enriched vacuolar proteins

69 proteins in the enriched group were annotated as vacuolar in the SGD, most of which are well-characterized vacuolar residents (Figure 4A). Their vacuolar localizations are well-established, which justifies the use of the annotations in iterative Group Analyses. These vacuolar proteins include 8 subunits of the vacuole V-type ATPase (A, C, D, H, E, a, d, and e) and canonical vacuolar proteases and phosphatises such as APE3, PRB1, CPS1, DAP2, and PHO8 as well as putative hydrolases ECM14, YBR139W, YNL115C and YNL217W.

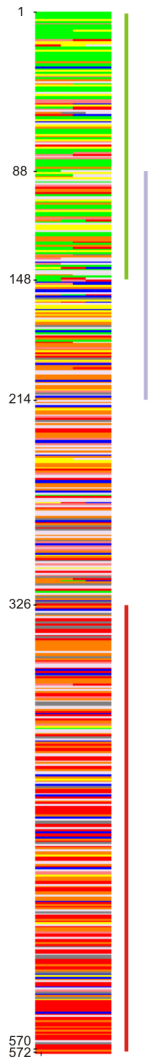


Figure 3. Enrichment clusters of proteins of different annotated localizations
All 572 identified proteins were ranked according to their iTRAQ ratios and db-iGA was applied to determine significant clusters of proteins with the same localization. The most enriched proteins are depicted at the top of the figure. Each horizontal bar in the column on the left represents a protein with (light yellow) or without (black) transmembrane domain(s) regardless of the subcellular localization. Each horizontal bar in column on the right represents a single protein colored according to the annotated subcellular localizations of a protein: green: vacuole; pink: plasma membrane; purple: ER, Golgi, endosome; orange: cytosol, cytoplasm; red: mitochondrial; gray: nucleus;

yellow: unknown; blue: others. The vertical bars indicate the range that shows a significant cluster of proteins from a particular subcellular localization as determined by db-iGA: green: vacuolar proteins, rank 1–148; light purple: endosome, rank 88-214; red: mitochondrion, range 326-570. The clustering of ER, Golgi nuclear, cytosolic and plasma membrane proteins was not significant.

The largest functional group of enriched vacuolar proteins contains 27 proteins with confirmed or predicted transport activity (Figure 4). Among them are 18 well-characterized vacuolar transport proteins, such as the glutathione S-conjugate transporter YCF1, the zinc transporter ZRC1, neutral amino acid transporters AVT1 and AVT3. Interestingly, we also detected AVT7, which is related to AVT1/3. Whereas the substrate specificity and the vacuolar localization of AVT1 and AVT3 were unambiguously demonstrated by immuno-fluorescence and transport activity measurements, AVT7 localization was ambiguous as it was found at the plasma membrane, or at the ER as GFP-tagged protein.[14,15] Our results clearly demonstrate a high enrichment of AVT7 in the vacuole.

The enriched proteins included 10 proteins annotated not only as vacuolar, but also bearing annotations for other subcellular localizations such as cytosol, mitochondria, endosome and plasma membrane (Figure 4A). A literature survey confirmed the vacuolar function of all 10 proteins, and interestingly, all of them are involved in membrane fusion, either directly, e.g. as members of the homotypic fusion and vacuole protein sorting (HOPS) complex, or indirectly by transmitting signals. Proteins involved in membrane fusion, trafficking, and targeting to the vacuole constitute the second largest functional group in the group of enriched proteins.

Depleted vacuolar proteins

The 424 depleted proteins should be regarded as non-vacuolar proteins (contaminants) and include for instance all 35 detected ribosomal proteins, the plasma membrane ATPase and the mitochondrial porin. However, this group also contained 8 proteins that were previously annotated as vacuolar. Among them, there are 4 well-characterized vacuolar proteins (VTC2, VTC3, VMA2, and PEP4). VTC proteins belong to the Vacuolar Transporter Chaperones, which are found at the ER and vacuole.[29] Upon induction of autophagy under nutrient limiting conditions, the VTC complex is recruited to vacuoles and concentrated at autophagic tubes of the

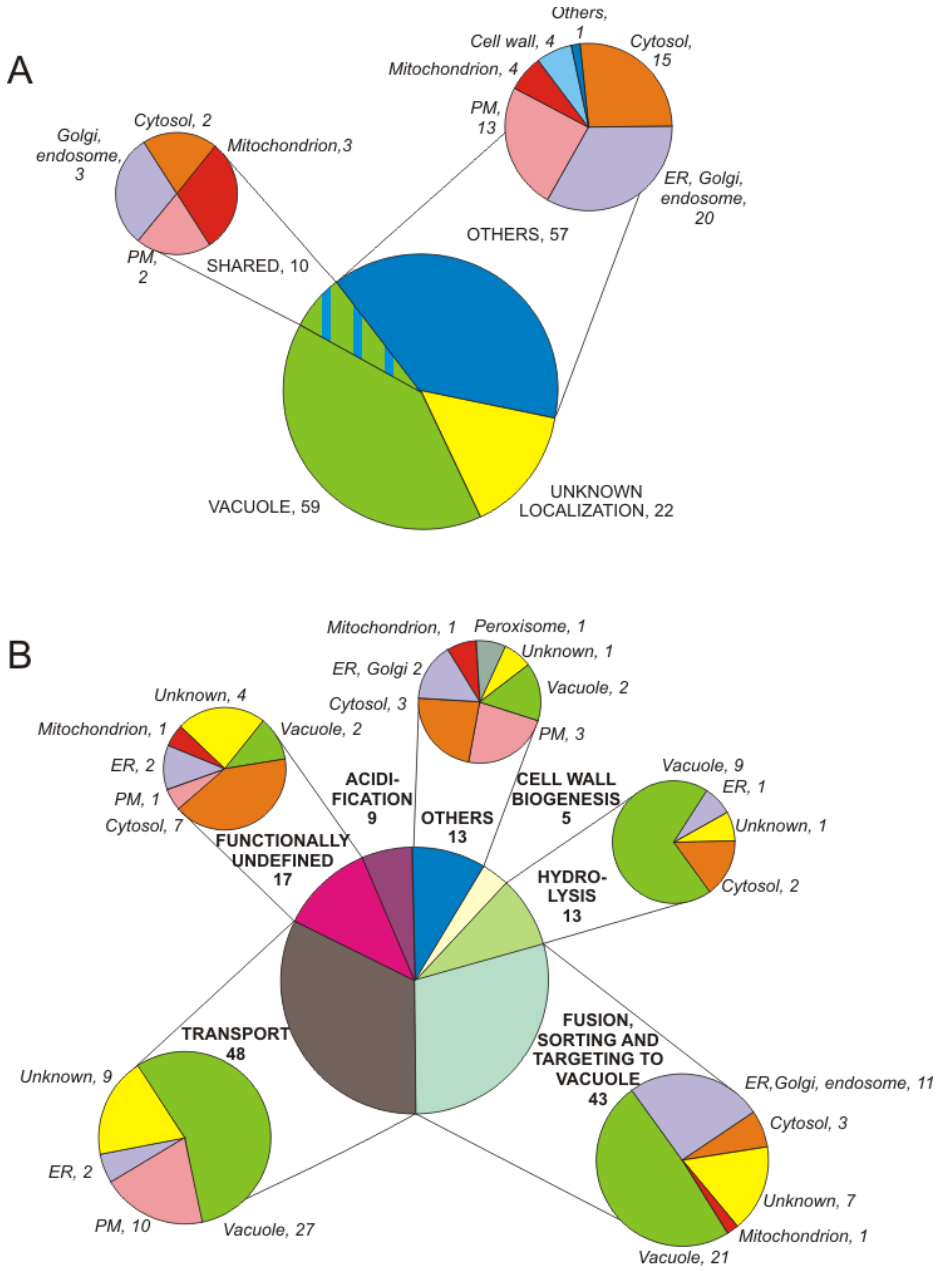


Figure 4. Grouping of enriched proteins

Panel A. Annotated localizations of enriched proteins. Db-iGA showed that 148 proteins were enriched together with known vacuolar proteins. 69 of these proteins had been assigned previously to the vacuole, including 10 proteins with multiple

localizations. Of the remaining proteins, 59 proteins had been annotated previously to other organelles than the vacuole. Colors as in Figure 3.

Panel B. Annotated functional categories of enriched proteins. The major functional group among the 148 enriched proteins represented 48 proteins that were known or were predicted to be involved in transport processes. 27 proteins from this group (54%) were known to be vacuolar. Twenty-two novel vacuolar proteins were distributed among the various functional groups (yellow segments).

membrane.[29] Interestingly, whereas VTC2 and VTC3 were depleted in our analysis, two other members of the vacuolar transporter chaperones, VTC1 and VTC4, were specifically enriched in the pure vacuolar membrane fraction suggesting distinct biological functions. The soluble subunit B (VMA2) of V-ATPase is loosely attached and may easily be lost during sample preparation, which could explain its apparent depletion.[30] The reason why PEP4 is depleted is not clear. However, it has been shown that PEP4 is secreted (instead of targeted to the vacuole) upon overexpression,[31] and the protein has also been found in highly purified mitochondria, indicating that the vacuole may not be the only or predominant destination of PEP4.

Four proteins (GRX7, SSA2, YKT6 and YKL077W) with shared localization in vacuoles and another organelle were depleted in our experiment. It is very well possible that GRX7 indeed is not a vacuolar protein, as a recent publication demonstrated a *cis*-Golgi localization of GRX7 (YBR014C), and its function is to catalyze glutathione-dependent reduction of disulfide bonds in oxidative stress conditions.[32] SSA2 is a member of heat shock protein 70 family and contributes to the transport of the vacuolar aminopeptidase APE1 and the cytosolic fructose-1,6-bisphosphatase FBP1 from the cytosol to the vacuole.[33-34] We could detect neither APE1 nor FBP1 in our vacuolar preparations, possibly because SSA2 was present at a different subcellular location, e.g. acting as chaperone in folding of newly synthesized cytosolic enzymes.[35-36] YKT6 is a protein with acyl-transferase activity, which is required for multiple protein targeting pathways to the vacuole participating in the *cis*-multi-SNARE complex.[37-38] This pathway includes the ER, where YKT6 is involved in ER to Golgi transport,[39] endosomes, where it is implicated in transport from Golgi to endosomes,[40] and the vacuole where it plays a role in homotypic fusion[37]. We could not identify any other components of the *cis*-multi-SNARE complex (VAM3, VAM7, NYV1) except for the v-SNARE protein VTI1, which was also depleted in our analysis. This indicates that the

cis-multi-SNARE complex was not located at the vacuolar membranes in our experiments. Finally, we could not confirm vacuolar localization of YKL077W, which was previously assigned to the vacuolar lumen based on a high-throughput localization study using GFP fusions,[15] and we suggest that it represents an incorrect annotation.

Undetected vacuolar proteins

In total we found 77 proteins that were annotated as vacuolar in the SGD. This covers 42% of all proteins annotated as vacuolar in the database. There are many possible reasons for not finding a large portion of proteins that were annotated as vacuolar. First, there is a group of proteins that have limited accessibility for digestion or identification by LC-MALDI due to their high hydrophobicity or small size, e.g. subunits c, c', c" of V-ATPase. Second, some proteins are of low abundance or may not be expressed under our experimental conditions. For instance, the vacuolar zinc transporter ZRT3 is expressed under zinc limiting conditions.[40] As yeast was grown on rich complete medium, it is unlikely that elements such as zinc ions were limiting for growth. Therefore, we assume that ZRT3 (among others) was not expressed to a sufficiently high level for detection. Third, because we aimed at the identification of membrane proteins, we stripped peripheral proteins from the vacuolar membranes using EDTA and high pH treatments. Many peripheral proteins, such as components of RAVE, SNARE and CCZ1-MON1 complexes, involved in vacuole biogenesis, homotypic vacuole fusion and protein targeting to the vacuole, may have been lost during sample preparation. Fourth, a number of database annotations are based solely on GFP-tagging studies, which could create localization artifacts. In those cases previous assignments as vacuolar may have been incorrect.

4.2 Enriched proteins with other localization

A group of 57 proteins with database annotations for various non-vacuolar localizations was enriched along with the known vacuolar proteins. It contained proteins of the ER, Golgi apparatus, endosomes, cytosol, plasma membrane, mitochondrion and cell wall (Figure 4A). The group of proteins with localizations other than the vacuole might be true vacuolar proteins with incomplete or incorrect annotation. For example, the putative protease YBR074W has no database annotation for localization, but is clearly enriched and is most likely a true vacuolar proteolytic enzyme. Similarly STV1, an isoform of VPH1 (subunit a of V-ATPase), was enriched in the pure vacuole fraction.

STV1 has been reported to be localized mainly in the Golgi and endosomes but our analysis shows that STV1 is also vacuolar. The two largest functional groups of enriched proteins without vacuolar database annotations are proteins involved in membrane fusion, trafficking and targeting to vacuoles and transport proteins.

5. Conclusion

We used the subtractive proteomics approach to identify organelle-specific proteins, in our case vacuolar membrane proteins. To deal with the problem of contaminants we used Localization of Organelle Proteins by Isotope Tagging (LOPIT). For analyzing the LOPIT data, we developed a statistical method (db-iGA) to determine the cluster of proteins enriched with vacuoles, as well as clusters of depleted proteins from different locations. As enrichment (and depletion) of certain classes of proteins is inherent to the LOPIT method, db-iGA proves to be a suitable method for analyzing such data. Our analysis yields insight in the dynamics of the vacuolar proteome (illustrated by proteins with multiple localizations), and in the function of vacuoles (degradation of proteins, novel putative transporters). We also found several proteins with erroneous localization annotations. Our work provides a solid basis for further characterization of vacuolar functions.

6. Acknowledgements

We thank Albert Sickmann, Fabrizia Fusetti, and Liesbeth Veenhoff for advise and helpful discussions in various stages of the project. The work was supported by The Netherlands Proteomics Centre (NPC), and The Netherlands Organisation for Scientific Research (NWO, vidi grant to DJS), and the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

7. References

1. Uchida, E., Ohsumi, Y., Anraku, Y., Purification and properties of H^+-translocating, Mg^{2+}-adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. ***J. Biol. Chem.*** 1985, 260, (2), 1090–1095.
2. Shirahama, K., Yazaki, Y., Sakano, K., Wada, Y., Ohsumi, Y., Vacuolar function in the phosphate homeostasis of the yeast *Saccharomyces cerevisiae*. ***Plant Cell Physiol.*** 1996, 37, (8), 1090–1093.
3. Indge, K. J., Polyphosphates of the yeast cell vacuole. ***J. Gen. Microbiol.*** 1968, 51, (3), 447-455.
4. Horst, M., Knecht, E. C., Schu, P. V., Import into and degradation of cytosolic proteins by isolated yeast vacuoles. ***Mol. Biol. Cell*** 1999, 10, (9), 2879–2889.
5. Rotin, D., Staub, O., Haguenauer-Tsapis, R., Ubiquitination and endocytosis of plasma membrane proteins: role of Nedd4/Rsp5p family of ubiquitin-protein ligases. ***J. Membr. Biol.*** 2000, 176, (1), 1–17.
6. Moeller, C. H., Thomson, W. W., Uptake of lipid bodies by the yeast vacuole involving areas of the tonoplast depleted of intramembranous particles. ***J. Ultrastruct. Res.*** 1979, 68, (1), 38–45.
7. Kim, I., Rodriguez-Enriquez, S., Lemasters, J. J., Selective degradation of mitochondria by mitophagy. ***Arch. Biochem. Biophys.*** 2007, 462, (2), 245–253.
8. Sakai, Y., Koller, A., Rangell, L. K., Keller, G. A., Subramani, S., Peroxisome degradation by microautophagy in *Pichia pastoris*: identification of specific steps and morphological intermediates. ***J. Cell Biol.*** 1998, 141, (3), 625–636.
9. Roberts, P., Moshitch-Moshkovitz, S., Kvam, E., O'Toole, E., Winey, M., Goldfarb, D. S., Piecemeal microautophagy of nucleus in *Saccharomyces cerevisiae*. ***Mol. Biol. Cell*** 2003, 14, (1), 129–141.
10. Wiemken, A., Schellenberg, M., Urech, K., Vacuoles: the sole compartments of digestive enzymes in yeast *Saccharomyces cerevisiae*. ***Arch. Microbiol.*** 1979, 123, (1), 23–35.
11. Sarry, J.-E., Chen, S., Collum, R. P., Liang, S., Peng, M., Lang, A., Naumann, B., Dzierszinski, F., Yuan, C.-X., Hippler, M., Rea, P. A., Analysis of the vacuolar luminal proteome of *Saccharomyces cerevisiae*. ***FEBS Journal*** 2007, 274, (16), 4287–4305.
12. Hurlimann, H. C., Stadler-Waibel, M., Werner, T. P., Freimoser, F. M., Pho91 Is a vacuolar phosphate transporter that regulates phosphate and polyphosphate metabolism in *Saccharomyces cerevisiae*. ***Mol. Biol. Cell*** 2007, 18, (11), 4438–4445.
13. Cagnac, O., Leterrier, M., Yeager, M., Blumwald, E., Identification and characterization of vnx1p, a novel type of vacuolar monovalent cation/H^+ antiporter of *Saccharomyces cerevisiae*. ***J. Biol. Chem.*** 2007, 282, (33), 24284–24293.
14. Russnak, R., Konczal, D., McIntire, S. L., A family of yeast proteins mediating bidirectional vacuolar amino acid transport. ***J. Biol. Chem.*** 2001, 276, 23849–23857.
15. Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., O'Shea, E. K., Global analysis of protein localization in budding yeast. ***Nature*** 2003, 425, (6959), 686–691.
16. Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., Weissman, J. S., Global analysis of protein expression in yeast. ***Nature*** 2003, 425, (6959), 737–741.

17. Miller, J. P., Lo, R. S., Ben-Hur, A., Desmarais, C., Stagljar, I., Noble, W. S., Fields, S., Large-scale identification of yeast integral membrane protein interactions. ***Proc. Natl. Acad. Sci. U. S. A.*** 2005, 102, (34), 12123–12128.
18. Kumar, A., Agarwal, S., Heyman, J., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K.-H., Miller, P., Gerstein, M., Roeder, S., Snyder, M., Subcellular localization of the yeast proteome. ***Genes Dev.*** 2002, 16, (6), 707–719.
19. Dunkley, T. P. J., Watson, R., Griffin, J. L., Dupree, P., Lilley, K. S., Localization of organelle proteins by isotope tagging (LOPIT). ***Mol. Cell. Proteomics*** 2004, 3, (11), 1128–1134.
20. Dunkley, T. P. J., Hester, S., Shadforth, I. P., Runions, J., Weimar, T., Hanton, S. L., Griffin, J. L., Bessant, C., Brandizzi, F., Hawes, C., Watson, R. B., Dupree, P., Lilley, K. S., Mapping the *Arabidopsis* organelle proteome. ***Proc. Natl. Acad. Sci. U. S. A.*** 2006, 103, (17), 6518–6523.
21. Breitling, R., Amtmann, A., Herzyk, P., Iterative Group Analysis (iGA): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. ***BMC Bioinformatics*** 2004, 5, (1), 34.
22. Lawson, J. E., Douglas, M. G., Separate genes encode functionally equivalent ADP/ATP carrier proteins in *Saccharomyces cerevisiae*. Isolation and analysis of AAC2. ***J. Biol. Chem.*** 1988, 263, (29), 14812–14818.
23. Kipper, J., Strambio-de-Castillia, C., Suprapto, A., Rout, M. P., Isolation of nuclear envelope from *Saccharomyces cerevisiae*. ***Methods Enzymol.*** 2002, Vol. Volume 351, pp 394–408.
24. Wiederhold, E., Veenhoff, L. M., Poolman, B., Slotboom, D. J. Proteomics of *Saccharomyces cerevisiae* organelles. ***Mol. Cell. Proteomics*** 2010, *9,* 431–445.
25. Ohsumi, Y., Anraku, Y., Active transport of basic amino acids driven by a proton motive force in vacuolar membrane vesicles of *Saccharomyces cerevisiae*. ***J. Biol. Chem.*** 1981, 256, (5), 2079–2082.
26. Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. ***Anal. Chem.*** 2002, 74, (20), 5383–5392.
27. Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. ***Anal. Chem.*** 2003, 75, (17), 4646–4658.
28. Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P., Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. ***FEBS Letters*** 2004, 573, (1–3), 83–92.
29. Uttenweiler, A., Schwarz, H., Neumann, H., Mayer, A., The vacuolar transporter chaperone (VTC) complex is required for microautophagy. ***Mol. Biol. Cell*** 2007, 18, (1), 166–175.
30. Bauerle, C., Ho, M. N., Lindorfer, M. A., Stevens, T. H., The *Saccharomyces cerevisiae* VMA6 gene encodes the 36-kDa subunit of the vacuolar H(+)-ATPase membrane sector. ***J. Biol. Chem.*** 1993, 268, (17), 12749–57.
31. Wolff, A., Din, N., Petersen, J., Vacuolar and extracellular maturation of *Saccharomyces cerevisiae* proteinase A. ***Yeast*** 12 1996, 12, (9), 823–832

32. Mesecke, N., Spang, A., Deponte, M., Herrmann, J. M., A novel group of glutaredoxins in the cis-Golgi critical for oxidative stress resistance. ***Mol. Biol. Cell*** 2008, 19, (6), 2673–2680.
33. Satyanarayana, C., Schröder-Köhne, S., Craig, E. A., Schu, P. V., Horst, M., Cytosolic Hsp70s are involved in the transport of aminopeptidase 1 from the cytoplasm into the vacuole. ***FEBS letters*** 2000, 470, (3), 232–238.
34. Brown, C. R., McCann, J. A., Chiang, H.-L., The heat shock protein Ssa2p is required for import of Fructose-1,6-Bisphosphatase into VID vesicles. ***J. Cell Biol.*** 2000, 150, (1), 65–76.
35. Kim, S., Schilke, B., Craig, E. A., Horwich, A. L., Folding in vivo of a newly translated yeast cytosolic enzyme is mediated by the SSA class of cytosolic yeast Hsp70 proteins. ***Proc. Natl. Acad. Sci. U. S. A.*** 1998, 95, 12860–12865.
36. Kweon, Y., Rothe, A., Conibear, E., Stevens, T. H., Ykt6p is a multifunctional yeast R-SNARE that is required for multiple membrane transport pathways to the vacuole. ***Mol. Biol. Cell*** 2003, 14, (5), 1868–1881.
37. Peplowska, K., Markgraf, D. F., Ostrowicz, C. W., Bange, G., Ungermann, C., The corvet tethering complex interacts with the yeast Rab5 homolog Vps21 and is involved in endo-lysosomal biogenesis. ***Dev. Cell*** 2007, 12, (5), 739–750.
38. Zhang, T., Hong, W., Ykt6 forms a SNARE complex with Syntaxin 5, GS28, and Bet1 and participates in a late stage in endoplasmic reticulum-Golgi transport. ***J. Biol. Chem.*** 2001, 276, (29), 27480–27487.
39. Lewis, M. J., Pelham, H. R. B, A new yeast endosomal SNARE related to mammalian Syntaxin 8. ***Traffic*** 2002, 3, (12), 922–929.
40. MacDiarmid, C. W., Gaither, L. A., Eide, D., Zinc transporters that regulate vacuolar zinc storage in *Saccharomyces cerevisiae*. ***EMBO J*** 2000, 19, (12), 2845–2855.

Chapter V

Statistical analysis of quantitative proteomics data derived from production of human CFTR in *Lactococcus lactis*

Parts of this chapter were published in *Mol. Cell. Proteomics*, DOI*:* 10.1074/mcp.M000052-MCP20 (2010)

Abstract

We used iTRAQ-based quantitative proteomics to investigate the physiological response of *Lactococcus lactis* upon the production of human CFTR. Protein abundances in membrane and soluble fractions were monitored as a function of induction time, both in CFTR expressing and non-expressing cells. 846 proteins were identified and quantified (35% of the predicted proteome), including 163 integral membrane proteins. We performed a rigorous statistical analysis of the quantification data using Rank Sum analysis-based p-values, which were in turn corrected for multiple comparisons using False Discovery Rate (FDR) analysis. With a FDR of 10%, we determined that in the membrane fraction the abundance of 147 proteins had significantly changed (70 up and 77 down) and in the soluble fraction 202 proteins (104 up and 98 down). Furthermore, using db-iGA we were able to identify in the membrane fraction enrichment of proteins belonging to the ribosomal cluster. The analyses led us to identify various stress responses pertinent to overexpression of human CFTR.

1. Introduction

One of the major challenges in quantitative proteomics is to discriminate between biologically significant and random changes in observed protein abundance. How does one establish a significance cut-off in a list of identified proteins with associated experimental quantification data? An often used strategy is to simply calculate a protein fold change ratio and then, using an arbitrary threshold, select significant changes. This can lead to being overcautious by using a very high threshold (large number of false negatives) or being optimistic with a low threshold (large number of false positives). In either case, it is not an efficient way of unlocking the information present in the data.

Statistical significance testing based p-values can solve this problem by assigning a probabilistic confidence level to the data. Quite simply, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that it was observed by chance (null hypothesis). The null hypothesis is often rejected and the observation accepted as significant when the p-value is less than 0.05, corresponding to a 5% chance of falsely rejecting the null hypothesis. Based on such analysis, proteins with a fold change associated p-value of less than 0.05

are generally accepted as significantly changed. This implies that the 5% chance of false positive is an acceptable risk. However, in a high-throughput environment this would be an inappropriate method to measure significance as the p-value is only statistically valid when a single score is computed. The reason being that instead of testing if one protein is significant using the p-value threshold, testing is done on potentially hundreds of different proteins. The massive scale of data produced in such experiments creates many opportunities to encounter the 5% threshold by chance. For example, for every 100 proteins tested, you would expect to find 5 with a p-value of 0.05 by chance. As such, in a high-throughput analysis such as for shotgun proteomics, the problem of multiple testing can lead to false-discovery of significantly expressed proteins.[1]

To correct for this type of an error, a multiple testing correction strategy needs to be employed. The goal is to adjust the p-values appropriately to account for multiple testing. One of the simplest approaches is the Bonferroni correction, which retains the prescribed family wise error rate (FWER), i.e. the probability of making one or more false discoveries when performing multiple comparisons. It does so by dividing the threshold by the total number of tests performed. For instance, if the goal is to use a FWER of 5% when testing 100 proteins, then instead of using a p-value of 0.05, the new threshold would be a p-value of 0.0005. In this manner, there is a 95% chance that none of the 100 proteins found to have a significant change in abundance are false positive. While this is an effective way of controlling the false positive rate, it comes at a cost of allowing too many false negatives. As a result, in studies with a little appetite for error, controlling the FWER using conservative methods such as Bonferroni correction is an accepted approach. Alternatively, in an exploratory study such as proteomics where significant results can often be re-tested using independent techniques, control of the false discovery rate (FDR) is often preferred. An FDR adjusted p-value (or q-value) of 0.05 implies that 5% of *significant tests*, instead of all tests as with normal p-values, can be expected to be false positives. As a result, FDR-based analysis provides the flexibility to perform multiple comparisons with the knowledge of how many false positives can be expected. The FDR approach estimates the adjusted p-values by trying to find the height where the p-values flatten out (Figure 1).

In this study, we relied on FDR-based analysis to investigate the use of *L. lactis* for the expression of the human cystic fibrosis transmembrane conductance regulator CFTR. We were able to express full length (1480 amino acids long) CFTR in the bacterial host, but the expression levels were too low to pursue structural studies, and expression was toxic to the cells. To understand this toxicity and to identify potential

remedies to improve expression levels, we investigated the physiological responses that were elicited in *L. lactis* upon CFTR expression by performing a global quantitative proteomics study.

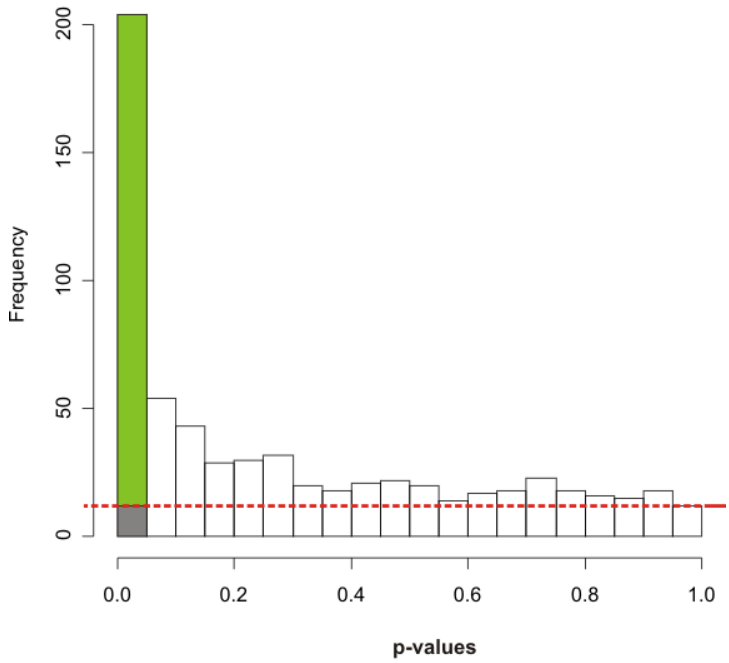


Figure 1. False Discovery Rate (FDR) analysis of p-values derived from *L. lactis* dataset

The x-axis shows histograms of p-values, each with a width of 0.05. The column with a p-value of less than 0.05 is shown in green. The horizontal red line is an approximation of where the p-values flatten out and the area under it represents the frequency of false positives. As such, the gray column is the expected frequency of false positives if a p-value of 0.05 is selected as the significance threshold.

2. Materials and Methods

2.1 Growth and preparation of samples

Growth in fermenters

L. lactis NZ9000 pNZ8048 (control strain) and *L. lactis* NZ9000 pNZncCFTR, containing the cloned human CTFR as described in Steen et al. were grown in 3 L pH and temperature-controlled bioreactors (Applikon) in M17 medium supplemented with

glucose (1%) and chloramphenicol (5 µg/mL). The temperature was set at 30°C and the pH was maintained at 6.5 during growth by addition of KOH. At an OD_{600} of 0.5 900 mL of the culture was removed and to the remaining culture Nisin A was added (1:5000 dilution of the supernatant of a culture of *L. lactis* NZ9700). After 1 hr and 4 hrs of induction 900 mL of the culture was collected. Cell were spun down (6,800 x g for 15 min, 4°C), and pellets were washed once with 10 mM potassium phosphate (KP_i) pH 7.5. The washed cell pellets were frozen in liquid nitrogen and stored at -80°C.

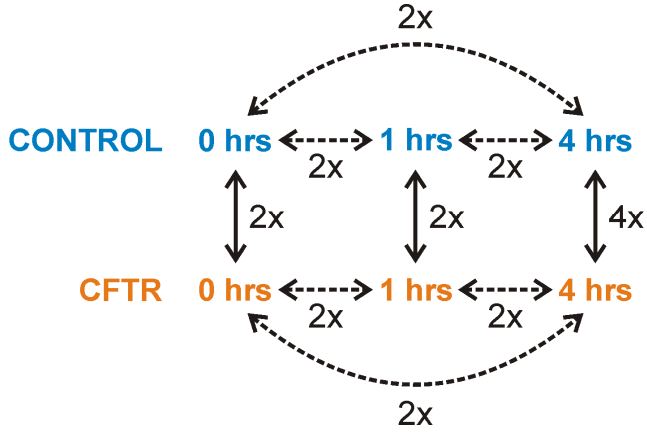


Figure 2. Schematic representation of a Master pool and outline of the meaningful comparisons of the relative protein abundances.
Each Master pool contained peptide samples representing three time points of the control strain and three time points of the CFTR-expression strain (0 h, 1 h, 4 h after induction expression). The expression levels of proteins could be followed as function of the time *within* the control and CFTR-expression strains (dashed lines). Furthermore, the relative changes in protein abundance could also be calculated *between* the control and the CFTR-expression strains at three the time points (solid lines). The comparison between the CFTR expressing strain and the control strain at the 4 h timepoint was repeated two more times in order to obtain 4 replicate values (4x).

Isolation of membrane and soluble protein fractions

The cell pellets were resuspended in 10 mM KP_i pH 7.5 at an OD_{600} of 50. To 6 mL of the suspension $MgCl_2$ was added (1 mM final concentration) and the cells were disrupted at 39 kPsi with a Constant Systems cell disrupter. The cells were passed through the disrupter cell twice. EDTA was added (15 mM) to the suspensions and they

were incubated on ice for 15 min. To remove non-broken cells the crude cell lysates were centrifuged for 15 min at 12,000 x g at 4ºC. The supernatant was carefully recovered and subsequently centrifuged at 267,000 x g for 15 min at 4°C. The supernatant, containing the soluble protein fraction was carefully pipetted off and stored at –80°C. Residual supernatant was completely removed from the membranes pellet. The membranes were washed once with 1 mL 10 mM KP_i containing 10% glycerol. The pellets were finally resuspended in 500 µL 10 mM KP_i, 10% glycerol and stored at –80°C. Protein concentrations were determined with the BCA kit (Pierce).

2.2 RP-LC and MALDI-TOF/TOF analysis

Peptides were trapped on a pre-column (300 µm x 5 mm, C18 PepMap300, LC Packing) and then separated on a C18 capillary column (C18 PepMap 300, 75 µm x 150 mm, 3 µm particle size, LC-Packing) mounted on the Dionex Ultraflex 3000 LC system (LC Packings, Amsterdam, The Netherlands). Mobile phase solutions contained A: 0.05% TFA; B: 0.05% TFA, 80% acetonitrile. Gradient conditions: equilibration of column, binding and washing of peptides was performed with 3% B, elution with 3 to 50% B in 60 min at a flow rate of 300 nL/min. The eluting peptides were mixed 1:4 with 2.2 mg/mL α-cyano-4-hydroxycinnamic acid matrix (LaserBio Labs, Sophia-Antipolis, France) and spotted directly onto a MALDI target (12 sec x 260 spots), using a Probot system (LC Packings, Amsterdam, The Netherlands). Peptides were analyzed with a 4800 Proteomics analyzer MALDI-TOF/TOF mass spectrometer (Applied Biosystems, Foster City, CA, USA).

The MALDI-TOF/TOF was operated in reflectron positive ionization mode in the m/z range 900-4000. The 15 most intense peaks above the signal-to-noise (S/N) threshold of 120 from each MS spectrum of odd-numbered RP-LC runs were selected for MS/MS fragmentation in the m/z range from 900 to 2000. The 10 most intense peaks above the S/N of 50 were selected from each MS spectrum of even-numbered RP-LC runs in the m/z range from 2000 to 4000. The MS/MS spectra were acquired using 2 kV acceleration voltage and air as collision gas at 5 x 10^{-7} Torr. The precursor mass transmission window was set to 300 (full width at half maximum, FWHM) for peptides in the m/z range of 900-2000, and to 200 (FWHM) in the range of 2000-4000 m/z. The peak-lists of the acquired MS/MS spectra were generated, using default settings and the S/N threshold of 10. The MS spectra were calibrated in the plate model mode, using 4700 calibration mixture (Applied Biosystems). MS/MS calibration of the

instrument was performed when required, using ACTH 18-39 (m/z = 2465.199) fragment ions.

2.3 Database search and criteria for protein identification

MS/MS peak-lists were extracted by the ProteinPilot software, version 2.0, using default parameters and were automatically submitted to a database search. All MS/MS spectra were analyzed using Mascot (Matrix Science, London, UK; version 2.0) and X!Tandem (www.thegpm.org; version 2007.01.01.1). Mascot and X!Tandem were set up to search a combined *L. lactis sp. cremoris* MG1363 database, allowing one missed cleavage of the digestion by trypsin. The database was created by combining forward and reversed entries of the *L. lactis* proteome (release version 31.08.07) and included sequences of porcine trypsin (NCBI accession: P00761), human keratins (P35908, P35527, P13645, NP_006112), chloramphenicol acetyltransferase (P00485), replication protein A (Q04138) and the human CFTR (NCBI accession: NP_000483) containing in total 4,902 protein entries. Mascot and X!Tandem searches were performed with a fragment ion mass tolerance of 0.30 Da and a parent ion tolerance of 200 ppm. MMTS modification of cysteine and Applied Biosystems 4-plex or 8-plexed iTRAQ quantitation chemistry of lysine and the N-terminus were specified in Mascot and X!Tandem as fixed modifications. Deamidation of asparagine and glutamine, oxidation of methionine and Applied Biosystems 4-plex or 8-plexed iTRAQ quantitation chemistry of tyrosine were specified in Mascot and X!Tandem as variable modifications.

Scaffold (version Scaffold-2_02_03, Proteome Software Inc., Portland, OR) was used to validate MS/MS-based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0% probability as specified by the Peptide Prophet algorithm (22). Protein identifications were accepted if they could be established at greater than 99.0% probability and contained at least 2 uniquely identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (23). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principle of parsimony. Those peptides were removed from the dataset when quantification was performed. The false positive rate was calculated by dividing 2 times the number of proteins identified in the reversed database by 4902, the sum of all proteins identified in forward and reversed versions of the database. In all measured

samples, no hits from the reversed database were detected, using the criteria described above.

2.4 Relative quantification of protein expression

The relative quantification was based on peptides that were chemically labeled with isobaric reagents, using the 4-plex or 8-plex iTRAQ technique. The quantification information was obtained from the peak areas of the reporter ions (m/z 112.2, 113.2, 114.2, 115.2, 116.2, 117.2, 118.2, 119.2 and 121.2). The peak areas were extracted from the MS/MS spectra by the ProteinPilot software using default settings as specified by the ProteinPilot for the 4800 MALDI instruments (Applied Biosystems). The peak areas were corrected for isotopic impurities by the ProteinPilot using the information provided by the manufacturer in the Certificate of Analysis for each iTRAQ batch. To select quantification data, those ratios were removed where the peak area of one reporter ion was below the signal-to-noise threshold of 10.

The global bias correction was performed for all identified peptides. The bias correction factor for a given iTRAQ ratio (*e.g.* 113/114) was calculated as the sum of all reporter peak areas in all measured spectra from one iTRAQ reagent (*e.g.* 114) divided by the sum of reporter peak areas of another reagent (*e.g.* 113). To obtain the bias-corrected peptide iTRAQ ratios, all measured ratios (in this example all 113/114 ratios) were multiplied by the correction factor. The bias-corrected peptide ratios of the same protein were weight-averaged and protein iTRAQ ratios were obtained according to the method utilized by the ProteinPilot software (Applied Biosystems). Peptides that matched to multiple proteins were excluded from quantification.

2.5 Statistical Analysis

To identify proteins with significantly changed abundances, we used FDR analysis on p-values computed by rank-based statistics. Rank Sum analysis was used to calculate p-values for the comparison of protein expression in CFTR expressing strain *vs.* the control strain at the 4 hrs time-point, where four independent replicates were available. Rank Sum is a non-parametric statistical method based on the Rank Product analysis (24, 25), which allows the data across biological replicates to be analyzed in a robust way. For the Rank Sum analysis the weighted protein ratios for each of the four replicate samples were calculated as described above and sorted in a descending order. Ranks were assigned to each protein, so that the protein with the highest ratio had rank

1, and the protein with the lowest ratio had a rank corresponding to the total number of identified proteins. To combine the protein ranks of all four measured replicates, the sum of ranks across replicates was calculated, sorted in descending order and ranked again. The p-value for each protein was calculated by comparing its rank sum with the result of 1000 permutations of the list using the RankProd package for R (26). The resulting p-values were then corrected for multiple testing using the adaptive FDR control method (27), giving the so-called q-values. This was done using the fdrtool R package (28). An FDR rate of 10% was used as the threshold for selecting proteins with significantly changed expression. The lists of proteins sorted by the RankSum were also used as input for double boundary iterative Group Analysis as described in chapter IV to analyze the ribosomal proteins.

3. Results and discussion

Production of sufficient amounts of well-folded membrane protein is a major bottleneck in membrane protein research. CFTR is no exception, and biochemical/biophysical studies on the protein are hampered by low yields of correctly folded and stable protein. Here, we have used the prokaryotic expression host *L. lactis* to express full-length human CFTR. Although the results are encouraging, the yields of CFTR were too low (<0. 1% of membrane protein) for functional/structural characterization. In addition, growth of the cells was severely compromised when expression of CFTR was induced, resulting in low biomass yield and indicating toxicity to the cell.

In an attempt to understand why the CFTR yields were low, and possibly to remedy the expression bottlenecks, quantitative proteomics was used to characterize the response of *L. lactis* to expressing CFTR in its plasma membrane. In the combined membrane and soluble fractions we identified and quantified a total of 846 proteins, representing 35% of the predicted *L. lactis* proteome. Among the identified proteins were 163 integral membrane proteins, which were strongly enriched in the isolated membrane fractions. The large number of identified proteins allows reliable analysis of the physiological responses of *L. lactis* to the expression of CFTR. To find proteins with significantly different abundances the RankSum and FDR algorithms were used as

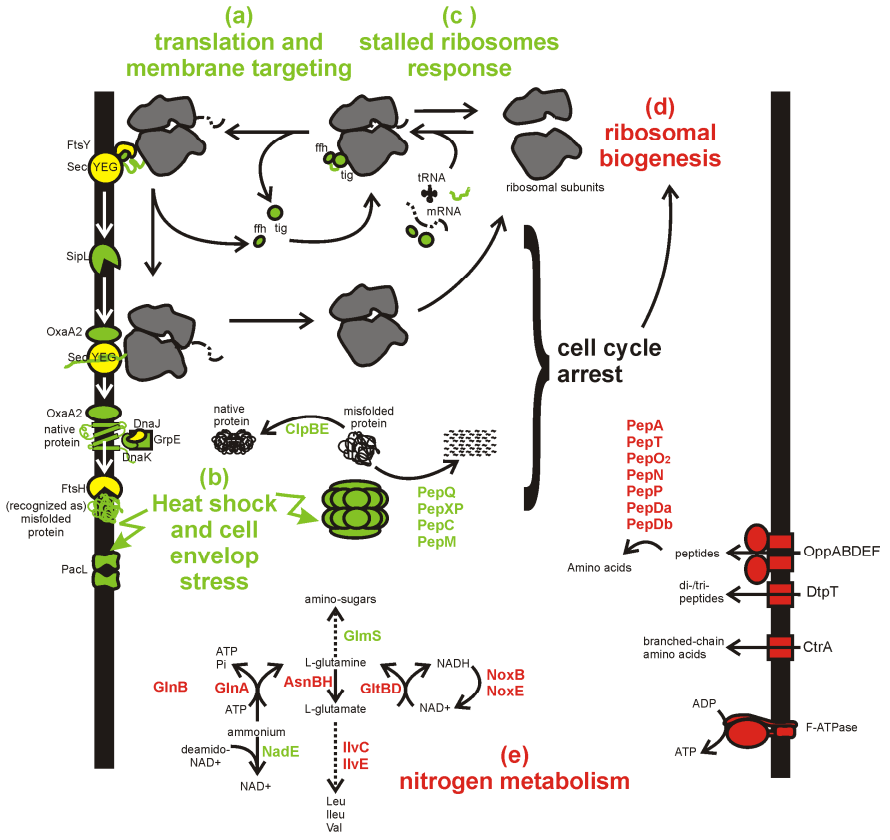


Figure 3. The physiological responses of the bacterium *Lactococcus lactis* to the production of the human CFTR.

Red and green colors indicate proteins that are lower and higher in abundance respectively in the CFTR expressing strain compared to the control strain at timepoint 4 hr after induction. Yellow proteins have the same abundance.

described in the methods section. At the 4 hrs time-point, 147 proteins had significantly changed in abundance in the membrane fraction (70 up and 77 down) and 202 proteins in the soluble fraction (104 up and 98 down) (FDR-corrected p-values <0.1). Expression of CFTR resulted in an increase in abundance of stress-related proteins (*e.g.* heat-shock and cell envelope stress), indicating the presence of misfolded proteins in the membrane. In contrast to the reported consequences of membrane protein overexpression in *E. coli*, there were no indications that the membrane protein insertion machinery (Sec) became overloaded upon CFTR production in *L. lactis*. Nutrients and ATP became limiting in the control cells as the culture entered the late exponential and

stationary growth phases but this did not happen in the CFTR expressing cells, which had stopped growing upon induction.

Furthermore, double-boundary iGA confirmed that almost all ribosomal subunits cluster among the proteins with the highest iTRAQ ratios in the membrane fraction but not in the soluble fraction. These results show that the distribution of ribosomes over the membrane and soluble fractions becomes different in the control and CFTR expressing cells as a function of induction time. The redistribution takes place predominantly in the control cells, where the fraction of membrane bound ribosomes decreases to a much larger extent than in soluble fraction. In the CFTR-expressing cells the distribution remains approximately the same. This finding is surprising, and shows that normal (control) *L. lactis* cells entering the late exponential/stationary growth phase specifically decrease the amounts of membrane bound ribosomes, perhaps indicating a lower need for integral membrane and secreted proteins. The major responses are summarized in Figure 3.

4. Conclusion

In this study, we performed a comprehensive statistical analysis based on false discovery rate (FDR) and double-boundary iterative Group (db-iGA) methods to investigate the effect of producing human CFTR in the prokaryotic expression host *Lactococcus lactis*. We used iTRAQ-based quantitative proteomics to find out why production of CFTR in *L. lactis* was problematic. Protein abundances in membrane and soluble fractions were monitored as a function of induction time, both in CFTR expression cells and in control cells that did not express CFTR. 846 proteins were identified and quantified (35% of the predicted proteome), including 163 integral membrane proteins. Using the FDR analysis, we found out that 147 proteins had significantly changed in abundance in the membrane fraction (70 up and 77 down) and 202 proteins in the soluble fraction (104 up and 98 down). Furthermore, we also discovered enrichment in the ribosomal cluster of proteins within the membrane fraction using db-iGA. This led to identifying various stress responses related to induction of human CFTR in *L.lactis*. The different stress responses elicited in *E. coli* and *L. lactis* upon membrane protein production indicate that different strategies are needed to overcome low expression yields and toxicity.

5. References

1. Hogg, C, Tanis, E. Probability and Statistical Inferences, 6th ed., Prentice-Hall: NJ, 2001.
2. Noble, W. S., How does multiple testing correction work? ***Nat. Biotech***. 2009, 27, (12), 1135–1137.
3. Benjamini, Y., Hochberg, Y., Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. ***J. R. Stat. Soc. Series B Stat. Methodol.*** 1995, 57, (1), 289–300.
4. Nguyen, D. V., On estimating the proportion of true null hypotheses for false discovery rate controlling procedures in exploratory DNA microarray studies. ***Comput. Stat. Data Anal.*** 2004, 47, (3), 611–637.
5. Steen, A., Wiederhold, E., Gandhi, T., Breitling, R., Slotboom, D. J., Physiological adaptation of the bacterium *Lactococcus lactis* in response to the production of human CFTR. ***Mol. Cell. Proteomics***. 2010, doi:10.1074/mcp.M000052–MCP201.
6. Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. ***Anal. Chem.*** 2002, 74, (20), 5383–5392.
7. Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P., Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. ***FEBS Letters*** 2004, 573, (1–3), 83–92.
8. Breitling, R., Herzyk, P., Rank–based methods as a non–parametric alternative of the T–statistic for the analysis of biological microarray data. ***J Bioinform Comput Biol.*** 2005, 3, (5), 1171–1189.
9. Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., Chory, J., RankProd: a bioconductor package for detecting differentially expressed genes in meta–analysis. ***Bioinformatics*** 2006**,** 22, (22), 2825–2827.
10. Benjamini, Y., Hochberg, Y., On the adaptive control of the false discovery fate in multiple testing with independent statistics. ***J. Educ. Behav. Stat.*** 2000, 25, 60–83.
11. Strimmer, K., A unified approach to false discovery rate estimation. ***BMC Bioinformatics*** 2008**,** 9, (1), 303.
12. Strimmer, K., FDRtool: a versatile R package for estimating local and tail area–based false discovery rates. ***Bioinformatics*** 2008**,** 24, (12), 1461–1462.
13. Breitling, R., Amtmann, A., Herzyk, P., Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. ***BMC Bioinformatics*** 2004, 5, (1), 34.
14. Wagner, S., Baars, L., Ytterberg, A. J., Klussmeier, A., Wagner, C. S., Nord, O., Nygren, P-Å., van Wijk, K. J., de Gier, J.–W., Consequences of membrane protein overexpression in *Escherichia coli*. ***Mol. Cell. Proteomics*** 2007**,** 6, (9), 1527–1550.

Chapter VI

Summary and perspectives

The analysis of the protein composition in complex biological samples inevitably starts with the collection of accurate and complete data sets. Protein composition analysis (proteomics) is nowadays almost always performed by shotgun proteomics using the analytical techniques of liquid chromatography (LC) coupled with tandem mass spectrometry (MS/MS). The LC-MS/MS technique is widely used not only for identification but also for quantification of proteins in complex samples. The ionization technique most commonly used in shotgun proteomics is electrospray ionization (ESI), but ionization by matrix-assisted laser desorption (MALDI) has certain advantages which make it a useful complementary technique to ESI. The relatively limited use of LC-MALDI-MS/MS for shotgun proteomics has had the consequence that several steps in the workflow are not fully efficient and optimized.

In this work, we developed methods to improve on the typical MALDI-based LC-MS/MS pipeline (chapter I). We approached this by investigating and addressing various bottlenecks in the pipeline and performing statistical analysis on the quantification data collected. This has led to the development of several algorithms and scripts. In chapter II, we present a new method, APECS, which lowers the redundancy, arising from broad peptide elution peaks, in the precursor selection process. This is done by exploiting the two-staged MS and MS/MS approach of the MALDI/TOF/TOF instrumentation, a step in the pipeline that offers many additional opportunities for data-dependent MS/MS analysis.

Chapter III revisits the traditional scoring system used for the identification of peptides by looking at the predictive power present in a peptide fragment's intensity. We underline that while there is indeed useful information present in the intensity of a peptide fragment, it is highly dependent on the sample preparation and instrument type. As such, a scoring model that makes use of this information will have to be adapted accordingly. Another concern lies in the observation that information available is often not sensitive enough to discriminate between two peptide sequences with similar mass-to-charge ratio (m/z). Nevertheless, elucidating the fragmentation pathways poses an exciting challenge with the potential to improve on the current m/z dependent protein identification methods.

As the field of MS-based proteomics matures, sophisticated statistical methods will play an increasingly important role in analyzing the high-throughput qualitative and quantitative data created from such experiments. Accordingly, chapter IV and V deal with the interpretation of quantitative results, using the DB-iGA and False Discovery

Rate-based analysis method for determining significantly changed expression patterns of proteins.

Besides the aforementioned results related to specific components of a proteomics pipeline, was there a single overarching lesson learnt during the work performed for this thesis?

Software remains a challenge

While the mass-spectrometry-based proteomics method has seen great instrument-related improvements in recent years, the proteomics pipeline is greatly constrained by the available software. Generally speaking, the current software solutions are not capable to fully harness the power offered by the instrument. For instance, as discussed in chapter II, data-dependent analysis can lead to a significant improvement of results obtained from a MALDI-TOF/TOF experiment, owing to its two-staged MS and MS/MS approach. However, software packages, especially the ones available from the original MS-instrument vendor, tend to treat LC-MS-based analysis as a linear pipeline. This means that attempts to perform data-driven, non-linear analysis often get stuck due to implementation-related challenges. Some of these challenges stem from lack of support for standard data formats and limited scope for creating customized software that would work seamlessly with the existing pipeline. These are exactly the kind of challenges that we experienced trying to use the methods created in this thesis. These problems are only aggravated by the inflexibility of the software cycle business model typically followed by instrument vendors. The version-based upgrades provided by them cannot keep pace with the method development being reported in the literature, even if one ignores the tendency to keep things "in-house". In this day and age, when consumer electronic product companies lead the market with their emphasis on *user-experience* and *openness*, the current level of accessibility and pace of improvement provided by the proprietary MS software feels obsolete. So, revisiting the question raised earlier, the lesson learned during this work is that MS software is lacking behind the available hardware and it will continue to remain so as long as proprietary software is being used.

The way forward

The key to improving the available software solutions lies in creating platforms which invite (software) developers from the proteomics community to create

applications, united by the tenets of the specific platform. Imagine a situation where the outcome of an MS-based method development research is an application that could be downloaded and used with the same ease by the community as a typical application for a smart phone. For this to happen, the community must move away from supporting standalone software suites that are often available for purchase from the instrument vendors. Instead, the community would stand to benefit in the long run if they back initiatives that facilitate community-based development. The end goal of such initiatives should be the creation of an application distribution platform similar to what has become a key ingredient in the telecommunications industry. Already, there have been notable initiatives in this direction from within the research community such as the Trans-Proteomics Pipeline (TPP) and The OpenMS Proteomics Pipeline (TOPP). For instance, TPP, encompassing all of the steps in a typical MS/MS pipeline, provides a robust platform for proteomics method development. It makes use of a standard data-format, mzXML, throughout its various applications and also provides scripts for the conversion of instrument-specific proprietary data-formats. Furthermore, development is supported by an extensive wiki-based documentation and mailing lists.

The primary advantage of the TPP-type of initiatives is that it encourages collaborative development. The tools created for them can also be easily shared with the other users. This leads to a more demand-driven growth of the platform when compared to the proprietary software where it is not always clear why certain "improvements" were made. In conclusion, the continued growth and acceptance of such platforms is required for bridging the gap between software and hardware and is the need of the hour.

Samenvatting in het Nederlands

Eiwitten spelen een uitermate belangrijke rol in alle levende organismen als de werkpaarden van de cel. De term proteomics, bedacht naar analogie van genomics, wordt vaak gedefinieerd als de uitgebreide kwantitatieve studie van eiwitexpressie en de fluctuaties daarin onder invloed van biologische veranderingen. Het doel van een typisch proteomics-experiment is om de eiwitten van een organisme te vergelijken die onder verschillende omstandigheden, zoals bijvoorbeeld temperatuur, een mutatie, beschikbaarheid van voedingsstoffen, ziek t.o.v. gezond, tot expressie komen. Het idee is het biologische systeem te begrijpen aan de hand van de verschillen tussen twee of meer toestanden, in het geval van proteomics aan de hand van de aanwezigheid, (kwalitatieve) toestand en hoeveelheden van eiwitten. Voordat een grondige kwantitatieve analyse van de verschillen kan worden uitgevoerd is het noodzakelijk dat alle onbekende eiwitten die aanwezig zijn in een biologisch monster worden geïdentificeerd. Shotgun proteomics vormt daarbij een essentieel gereedschap in high-throughput analyse van eiwitten in complexe biologische monsters. Precieze identificatie van peptides afkomstig van vloeistofchromatografie-analyse gekoppeld aan tandem-massaspectrometrie vormt de basis van dergelijke analyses. Het werk beschreven in dit proefschrift behelst de verbetering van de prestaties van een op LC-MALDI gebaseerd proteomics werkschema, waarbij de focus ligt op het verbeteren van de specifieke componenten hiervan. Dit heeft geresulteerd in verschillende nieuwe algoritmes en softwarematige oplossingen.

Acknowledgements

The trials and tribulations of PhD can overwhelm even the most collected of a person. Given a chance, the feelings of insecurity and guilt are always right around the corner waiting to entrap you within its sinister web. In such an environment, the scientific and social support system becomes a crucial ingredient for the success of any student. Here, I would like to acknowledge the many faces that have been part of my support system.

First of all, I will thank my daily supervisor and co-promoter, *Hjalmar Permentier*. I find that the parallels between my time as a student and his personal and professional life make for an engaging story: he became a husband at the beginning of my studies and a father towards its end. In between, he also found much success in his career. Hjalmar has a great wealth of knowledge, but with none of the trappings of pride. In addition to his top-notch expertise with *all things* MS, he also makes for a pretty decent part-time (bio)informatician, spaghetti-code notwithstanding.

I also thank *Rainer Breitling*, who played a very important role as the sole bioinformatician amongst my promoters. Throughout my project, I relied upon him to provide a candid assessment of my progress. Through our discussions, he always brought clarity and direction to my research. As such, I have fond memories of leaving Haren inspired and hopeful.

Finally, I would like to thank *Bert Poolman*, also my promoter. He has always been fair and made things easier in more ways than one. He has always been a source of inspiration through his passion for science. Specifically, his input during the writing of this book was invaluable.

Besides my promoters, there are others who played an important role in my scientific growth. *Dirk Slotboom* recruited me to perform data-analysis. This was well-timed as it came during the "dark-ages" of my PhD. Another person who helped my scientific confidence by consistently showing an interest in my work was *Liesbeth Veenhoff*. It was a pleasure doing some programming for her – even though it did not quite work out. She also helped by reviewing my manuscripts and the introduction

chapter of this very book! I also shared many fruitful discussions with *Fabrizia Fusetti*, the *doorkeeper* of all the MS instruments in our group. It was always exciting brain-storming with her as she is always full of ideas.

On the administrative side of things, I am deeply grateful to *Karlien Groothoff* for all of her help during my struggles with all things administrative.

Amongst my peers, *Elena Wiederhold* played a crucial role in my scientific development. She took the role of holding my hand and walking me through the world of MS-based proteomics. We also shared several fruitful collaborations which resulted into three publications. Even after moving to Zurich, she remained an important resource through the wonders of instant messaging. While I will always harbor a certain amount of misgiving towards her for starting the *typical Tejas* cliché, she will always have my heartfelt thanks for being a great mentor and friend.

Perhaps because I never stepped into the lab, it seemed as if our group had the friendliest bunch of scientists. While it would be difficult to list everyone by their names, I would like to mention at least some of the past and present members of the group: *Mr. Groeneveld* for being ever so helpful, *Ronnie* for being a proponent of warm lunches, beards, and positive thinking, *Ravi* for keeping his desk stocked with cookies, *Josy* for looking the other way when I raided the said cookies (and sometimes joining in), *Jacek* for keeping things colorful and inoculating me of any superstitions related to flying, *Justyna* for caring more about my moose pillow than myself, *Gea* for all the eggs, spinazie, and Indian food, *Wim* for all his jokes, *Astri* for being downright crazy, *Inga* for all those light and healthy desserts, *Faizah* for being even more clueless than me when it came to group gossips, *Karin* for teaching me settlers, *Jan-Peter* for at least one inappropriate joke during the lunch, *Anton* for being a good sport despite being the target of the said jokes, *Marysia* for all those delicious cakes, *Pranav* for asking small questions that often led to long answers, *Gemma* for being an "angel", *Ria* for always looking after everything, and *Armagan* for never failing to smile when passing by in the corridors.

Outside of work, I spent a large amount of my time interacting with the Winschoterkade crowd. *Radek Šimík*, *Jorge Tendeiro, and Rei Mondein Tendeiro* became close friends with whom I shared many laughs. *Inken* still amazes me with her willingness to help just about everyone. *Sushma* came for a few months like a whirlwind and shook me out of my slumber. A shout-out goes to *Berfu*, *Ebru*, *Katherina*, *Linda* and *Deepa*. While not inhabitants of the famous house, it would be appropriate to thank here my dear friends *Zhenya*, *Navi*, and *Sasha* for making me a part of their family. I will always

treasure the opportunity I have had to watch Sasha grow. Also, *Aysa Arylova,* another good friend.

I would also like to thank both *Bruno* and *Christin* for taking me under their wings when I first moved to Groningen. Their friendship was one of the reasons that I was willing to give this place a shot. *Kicki*, the past several years has seen our friendship grow stronger and you know how close you are to my heart. While leaving Sweden, I was not sure if we would manage to stay in touch, this time there are no such lingering doubts. Also, heartfelt thanks go out to *Darima* and *Yoana*, two more close friends from Chalmers. I would also like to mention *Nicolas* here.

Going further away from present time and place, I would like to thank few of the people who have played an important role in my journey to this day: *Stefanie "First Elf"* for being one of my greatest ally and introducing me to Chalmers, my land family from Gothenburg (*Agneta*, and *Arne Billestedt*) who made my time there so special, my Shyamal family (*Darshana* aunty, *Pranav* kaka, *Salomi* and *Sahil*) for always being a source of support, all of my colleagues from *Electronic Arts* (*Bonnie, James, Aira, and others*) who helped me learn to tap into my creative side, my professors from MSU (*Sarah Kruse*, *Dr. Christophe Veltsos*, *Dr. Timothy Secott*, *Dr. Ann Quade,* and *Dr. Dean Kelly*) who exposed me to the world of academia, my high-school chemistry teacher (*Brenda Ramin*), *Joy Cogan* who helped me come up with the idea for the cover of this very book, *Jos Boumans* and *Theresa Schnedier* for being there when the stars were dim, and *Jessica Matelski* for being a gangster's girlfriend.

Stemming from a family where education was *always* given one of the highest priorities, it is of no surprise that mine is crowded with engineers, PhDs, and medical doctors (in one person's case, all three), with a spattering of MBAs. As a result, I have never suffered from a lack of career role-models or advice. For this, I would like to thank every single overachieving member of the greater *Gandhi* and *Parikh* family (this also includes you *Ringwalas*, *Bhatts*, and others). Most of all, I would like to thank my parents, *Rashmi* and *Paresh Gandhi*, who have made so many sacrifices on my behalf, my brother and sister-in-law, *Ashish* and *Lori Gandhi*, who have always supported me, and my nephews and niece, *Ethan*, *Luka*, and *Priya*, who give the best hugs.

T. P. Gandhi

Publications

- ✓ **Gandhi, T.**, Fusetti, F., Wiederhold, E., Breitling, R., Poolman, B., Permentier, H.P. (2010). Apex peptide elution chain selection: a new strategy for selecting precursors in 2D-LC-MALDI-TOF/TOF experiments on complex biological samples, *J Proteome Res* 9, 5922-5928.

- ✓ Steen, A., Wiederhold, E., **Gandhi, T.**, Breitling, R., Slotboom, D.J. (2010). Physiological response of the bacterium *Lactococcus lactis* to production of human CFTR, *Mol Cell Proteomics*, DOI*:* 10.1074/mcp.M000052-MCP20.

- ✓ Wiederhold, E., **Gandhi, T.**, Permentier, H.P., Breitling, R., Poolman, B., Slotboom, D.J. (2009). The yeast vacuolar membrane proteome, *Mol Cel Proteomics* 8, 380-392.

Epilogue

It was a clear December night in Groningen as I biked along my favorite path next to the canal. Looking up to catch a glance of the moon, I was surprised to discover a large flock of migrating birds directly above me. With a gentle breeze brushing past me, I stretched my hands wide with an open palm, hardly noticing as my body balanced itself to negotiate a sharp turn. Then, closing my eyes for a moment, it felt as if I was tens of meters above the land along with my feathery friends.

Since that uncharacteristic autumn day, the past two years had seen an immense improvement in my strength and balance. Having learned to step out of my comfort zone, there was a world of new experiences waiting. No longer afraid of stumbling, my confidence grew as I challenged myself with more difficult tasks; the sharper the turn of the road, more thrilling the undertaking. But, as the last of the flock disappeared in the night, I was left with the sudden awareness that the day was not far when I too will have to set out to discover new roads with an entirely new set of twists and turns. Deciding that nostalgia would have to wait, I grinned as I went down a familiar path, one more time, chasing the wind with my Batavus.